DOCUMENT RESUME

ED 109 138 TM 004 390

AUTHOR	Durost, Walter N.; Hodges, Richard B., Jr.
TITLE	A Study of Test-Taking Behavior for Two Independent Samples of Pupils at Grade Four with Special Emphasis on Guessing.
INSTITUTION	Test Service and Advisement Center, Lee, N.H.
PUB DATE	[Nov 74]
NOTE	147p.
EDRS PRICE	MF-$0.76 HC-$6.97 PLUS POSTAGE
DESCRIPTORS	Academic Achievement; Achievement Tests; Disadvantaged Youth; Elementary Education; Grade 4; *Guessing (Tests); Intelligence Tests; Item Analysis; *Multiple Choice Tests; Response Style (Tests); Scores; Standardized Tests; *Testing Problems; *Test Wiseness; *Timed Tests
IDENTIFIERS	Elementary Secondary Education Act Title I; ESEA Title I; *Measurement of Change; New Hampshire

ABSTRACT

Data available from the Fall and Spring administration of the Stanford Achievement Test: Intermediate I: Form X and the Otis-Lennon Mental Ability Test for all Title I pupils in the State of New Hampshire plus a random sample for the entire state, made possible the item-by-item comparison of Fall and Spring performance on the same groups of children. Analysis of data was limited to one grade and five tests out of the complete battery. The purpose of the exploratory data analysis was to examine: relative gains by item, proportion of items attempted which were answered correctly, guessing, the prediction of a total score from a time-limited score on the basis of item analysis data available for the State of New Hampshire, and other incidental findings that might provide leads in constructing improved tests, especially for the measurement of Fall and Spring changes in educational achievement and particularly for the disadvantaged pupils as represented by the Title I selection within New Hampshire. Ways are now being sought to guarantee that children will work for the allowed time at the level effective for them, while at the same time predicting their score on an untimed test of "n" items. (Author/BJG)

****
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
****

ED109178

A STUDY OF TEST-TAKING BEHAVIOR FOR TWO INDEPENDENT SAMPLES OF PUPILS AT GRADE FOUR WITH SPECIAL EMPHASIS ON GUESSING

* * *

U S DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

TM 004 390

Prepared Jointly By
WALTER N. DUROST, Director
Test Service and Advisement Center
Lee, New Hampshire

and

RICHARD B. HODGES, JR., Coordinator-Director
Title I, ESEA, New Hampshire State Department of Education

PREFACE AND ACKNOWLEDGMENTS

This study did not follow any conventional research design. It was exploratory, unconventional, and uncommited to any established Truths about tests and testing. It grew out of forty years of constant involvement in testing, but did not hold any part of present practices in test construction and use as sacrosanct.

Its purposes were not destructive but hopefully creative, with the constant goal of learning how to make better tests. The roots of the study go back twenty years or more, when I became profoundly dissatisfied with grade equivalents as a means of interpretation of tests for individuals, plus my feeling of frustration over the waste of invaluable data insufficiently used to improve education, but particularly classroom instruction.

It was also particularly motivated by a growing dissatisfaction with the run-of-the-mill multiple choice-type item, which in this writer's opinion is a fraud and a sham unless concrete steps are taken to bring the RIGHT choice of answer in line with the actual existence of the knowledge it is intended to reflect.

Many people have contributed to this report in diverse ways. The U. S. Office of Education, Federal/State Developmental Staff, of which Dr. Richard M. Jaeger was the Acting Director, encouraged the study and also was instrumental in obtaining some funds to help in defraying part of the original expense. Dr. Jaeger gave further assistance with special regard to the intercorrelations of Rights and Attempts, using computer facilities available to him.

This phase of the study started new lines of thinking about further data analysis because the correlation matrix was so unusual. The variables were not linearly related, but the computer did not know that!

Mr. Richard B. Hodges, Jr., Coordinator-Director, Title I, ESEA, New Hampshire State Department of Education, is appropriately named as co-author because of his continued professional and financial support.

While Mr. Hodges did not participate directly in the writing of this final report, he has read it page-by-page and in numerous conferences has made many suggestions for revisions and concerning the interpretation of the data. Without his personal assistance, going back to the conception of the project, it would never have reached the published form.

Many of the problems involved in the collection of the original data were shared with the author while he was a consultant to Title I in 1968-1970 and responsible for an original report to the New Hampshire State Department of Education on the state testing program of these years. Considerations growing out of this experience suggested further item analysis, using the answer sheets to obtain item-to-item comparisons.

The significance of this idea, while it was apparent to both of us from the beginning, has grown to totally unexpected proportions as the analysis and writing of the report continued.

To my secretary, Mrs. Lois Mikoloski, goes my unbounded gratitude for invaluable day-by-day support and for many very perspicacious suggestions and comments.

Walter N. Durost
November 1974

ANSWER SHEET STUDY - GRADE 4 - 1969-70

TABLE OF CONTENTS

Page
PREFACE and ACKNOWLEDGMENTS i

PART I HISTORICAL BASIS AND PHILOSOPHICAL PERSPECTIVE FOR THE PRESENT STUDY
  Introduction I- 1
  Some Pertinent Facts in Regard to the Stanford Achievement Test: Intermediate I Battery: Form X I-15

PART II ANALYSIS OF DATA FOR THE RANDOM SAMPLE - Fall 1969 and Spring 1970
  Introduction II- 1
  Distribution Characteristics II- 1
  Difficulty Characteristics of Each of the Items in the Five Tests II- 8
  Analysis of Pupil Responses by Category II-15
  Some Characteristics of Standardized Tests Relevant to the Relationship Between Rights and Attempts II-22
  The Guessing Index or The Rights/Attempts Ratio II-28
  Use of Predicted Score to Determine Extent of Guessing II-34
  Evaluating Pupil Performance from the Pupil Rosters II-41
  Sample Cases for Illustrative Purposes II-45
  Sample A II-47
  Sample B II-52
  Sample C II-57

PART III COMPARISON OF TITLE I CASES WITH RANDOM SAMPLE ON ALL ESSENTIAL VARIABLES
  Introduction III- 1
  The Title I Data Compared With Random Sample III- 5
  Analyses and Comparison of Item Difficulty Information: Title I III-19
  Analyzing the Fall-Spring Pupil Results by Category (Review) III-26
  Rights versus Attempts for Title I Compared to Random Sample III-31
  The R/A Ratio as Applied to Title I III-37
  Use of Item Difficulties for Predicting Total Score on a Test III-39

SUMMARY and CONCLUSIONS

LIST OF APPENDICES

Appendix A Frequency Distributions, Cumulative Percent Distributions, and Stanines Plus Histograms Showing Shape of the Raw Score Distribution Graphically

  A1 Random Sample - Paragraph Meaning - Fall 1969 1
  A2 Random Sample - Paragraph Meaning - Spring 1970 2
  A3 Random Sample - Arithmetic Concepts - Fall 1969 3
  A4 Random Sample - Arithmetic Concepts - Spring 1970 4
  A5 Random Sample - Arithmetic Applications - Fall 1969 5
  A6 Random Sample - Arithmetic Applications - Spring 1970 6

Appendix B Superimposed Correlation Plots for Predicted and Actual Scores Separately for Those Who Attempted All Items and for Those Who Did Not

  B1 Random Sample - Word Meaning - Boys 7
  B2 Random Sample - Word Meaning - Girls 8
  B3 Random Sample - Arithmetic Computation - Boys 9
  B4 Random Sample - Arithmetic Computation - Girls 10
  B5 Title I - Word Meaning - Boys 11
  B6 Title I - Word Meaning - Girls 12
  B7 Title I - Arithmetic Computation - Boys 13
  B8 Title I - Arithmetic Computation - Girls 14

Appendix C Actual versus Predicted Scores Correlations and Attempted-All versus Did-NOT-Attempt-All Comparisons for Selected Statistics - Random Sample - Arithmetic Concepts and Arithmetic Applications 15

LIST OF TABLES

Table		Page
I- 1	Percentiles and Grade Equivalents Corresponding to Selected Percentile Ranks	I-10
I- 2	Composition of Stanford Achievement Test: Intermediate I: Form X and Composition of Otis-Lennon Mental Ability Test: Elementary II: Form J	I-16
II- 1	Item Difficulty Values - Percent Answering Each Item Right, Wrong or Omit and Rights/Attempts for Each Item - Random Sample	II- 9
II- 2	Correlations in Raw Scores Between Otis-Lennon and Selected Stanford Tests - Random Sample	II-12
II- 3	Number and Mean of Pupil-Item Responses by Consistency Category - Random Sample	II-16
II- 4	Number and Percent of Pupil-Item Responses by Consistency Category - Random Sample	II-19
II- 5	Analysis of Categories "WR+OR" and "RW+RO" - Random Sample	II-20
II- 6	Attempts versus Rights - 100-case Random Sample	II-24
II- 7	Attempts versus Rights - 567-case Random Sample	II-25
II- 8	Distribution of Attempts - Random Sample	II-26
II- 9	Distribution of Right Responses for Students Who Attempted All Items - Random Sample	II-27
II-10	Rights/Attempts - Means and Standard Deviations - Random Sample	II-30
II-11	Actual versus Predicted Scores - Correlations - Random Sample	II-37
II-12	Attempted-All versus Did-NOT-Attempt-All Comparisons for Selected Statistics - Random Sample	II-38
III- 1	Raw Score Means, Medians, and Standard Deviations - Title I	III-11
III- 2	Percent Answering Each Item Right, Wrong, or Omit and Rights/Attempts for Each Item - Title I	III-20
III- 3	A Comparison of the Percent Passing Each Item for the Random Sample Population and the Title I Population	III-23
III- 4	Number and Percent of Pupil-Item Responses by Consistency Category - Title I	III-27
III- 5	Number and Mean of Pupil-Item Responses by Consistency Category - Title I	III-28
III- 6	Analysis of Categories "WR+OR" and "RW+RO" - Title I	III-29
III- 7	Attempts versus Rights - Title I	III-32
III- 8	Distribution of Attempts - Title I	III-34
III- 9	Distribution of Right Responses for Students Who Attempted All Items - Title I	III-35
III-10	Attempted-All versus Did-NOT-Attempt-All Comparisons for Selected Statistics - Title I	III-40

LIST OF TABLES (Cont'd)

Table Page

III-11 Actual versus Predicted Scores - Correlations - Title I III-42

III-12 Distribution of Corrected Scores Which Would Have Been Reduced to Zero or Negative Values - Two Tests Fall and Spring - Boys - Title I III-43

LIST OF FIGURES

Figure

I- 1 Sample Roster Page of Pupil-Item Responses by Alternative Numbers I-11

Frequency Distributions, Cumulative Percent Distributions, and Stanines Plus Histograms Showing Shape of the Raw Score Distribution Graphically - Random Sample:

II- 1 Word Meaning - Fall II- 2
II- 2 Word Meaning - Spring II- 3
II- 3 Arithmetic Computation - Fall II- 6
II- 4 Arithmetic Computation - Spring II- 7

II- 5 Sample Pupil Record Card II-28

II- 6 Rights/Attempts Stanine Distributions - Random Sample - Fall II-32

II- 7 Rights/Attempts Stanine Distributions - Random Sample - Spring II-33

II- 8 Roster Page by Alternative Numbers II-42

II- 9 Roster Page by Right, Wrong, Omit Mode II-43

Frequency Distributions, Cumulative Percent Distributions, and Stanines Plus Histograms Showing Shape of the Raw Score Distribution Graphically - Title I:

III- 1 Word Meaning - Fall III- 6
III- 2 Word Meaning - Spring III- 7
III- 3 Arithmetic Computation - Fall III- 9
III- 4 Arithmetic Computation - Spring III-10
III- 5 Arithmetic Applications - Fall III-13
III- 6 Arithmetic Applications - Spring III-14
III- 7 Paragraph Meaning - Fall III-15
III- 8 Paragraph Meaning - Spring III-16
III- 9 Arithmetic Concepts - Fall III-17
III-10 Arithmetic Concepts - Spring III-18

LIST OF CHARTS

Chart

I- 1 Demonstration Bivariate Showing One-to-One Relationship Between Corrected and Uncorrected Scores When All Items are Attempted I- 3

I- 2 Normal Percentile Chart - Distribution of Otis-Lennon IQs for Tested Random Sample and Total State - Fall 1969 - Grade 4 I- 6

II- 1 Normal Percentile Chart - Distribution of Rights/Attempts - Random Sample II-29

III- 1 Normal Percentile Chart - Distribution of Rights/Attempts - Title I III-38

PART I

HISTORICAL BASIS AND PHILOSOPHICAL PERSPECTIVE FOR THE PRESENT STUDY

INTRODUCTION

At the beginning, this study started out to be strictly an empirical analysis of available data to obtain some better insight into the psychology and mechanics of guessing. As the data began to unfold, however, it was apparent that far more was involved than guessing alone. Indeed, the entire fabric of test-utilization behavior was in question. Many questions about the real reasons for testing under different circumstances very quickly came to mind.

New emphasis on item analysis as a means of interpreting test results reenforced the idea that the answers to multiple choice questions were too casually being equated to actual work-sample performance. It was clear that the functioning of any item was more closely related to its apparent ease or difficulty for the test-taking group than we had sensed in the past.

It was also realized that some subject matter was far more readily adaptable to objective-type test items than other types of content. Finally, a little thought convinced this writer that modification of the item types being used could practically eliminate the guessing factor, without making the test impossible to score electronically, while yielding substantial amounts of data not now being obtained.

However, the first priority in the analysis of the available data is still the investigation of GUESSING behavior, how it can best be eliminated or how it can be counteracted, and how the general attitude toward testing on the part of teachers and pupils can be improved with the consequent improvement of the educational process - especially with educationally handicapped children.

Guessing

The study of the effect of guessing in objective-type examinations has been a concern of the test-makers and publishers almost from the beginning of the effort to construct such tests.

For example, the first achievement test battery to be published and widely used throughout this country was the Stanford Achievement Test, Forms A and B, Copyright 1923. These tests, covering grades 2 - 8, included a wide variety of items, all rather steeply graded with respect to difficulty.

The Primary Battery of this series consisted of certain items from the beginning part of the Advanced Battery, so there was no actual differentiation in content between the Primary Battery and the Advanced Battery. In other words, the so-called Primary Battery was not truly a primary grades battery in the current sense of the word.

In several of the subtests in this pioneering test series, items which were of the multiple choice-type were used, and where this was the case a correction for guessing was indicated in the scoring directions - although the Manual of Directions contained no specific instructions about the effect of guessing or not guessing and no instructions to guess or not to guess.

More specifically, the score for the Reading: Sentence Meaning Test was indicated to be number right minus number wrong. The pupil directions for this test read as follows:

"Read the first sentence at the top of the page. It says: 'Can dogs bark? Yes No.' The right answer is 'Yes', so the word Yes has a line under it.

"Look at the second sentence (slowly). 'Does a cat have six legs? Yes No.' This time the correct answer is 'No', so the word No has a line under it. Now you must read each question on this page and draw a line under the right answer. Ready? Go."

Note that no indication was given to the students that the score would be rights minus wrongs, as stipulated in the scoring directions.

A similar correction for guessing was employed in Test 6: Nature Study and Science which used three-choice multiple choice questions, the directions for scoring saying simply that the score was number right minus one-half the number wrong. A comparable correction was used in Test 7: History and Literature. In Test 8: Language Usage, the score once more was number right minus number wrong. All other tests in this battery were of such a type as not to allow for a correction for guessing.

Dr. Giles M. Ruch was the junior author of the Stanford Achievement Test, along with Lewis M. Terman and Truman L. Kelley - both very well known educational psychologists. Giles M. Ruch seems to be the one in this authorship team who was particularly interested in the problem of guessing and its effect on reliability and validity.

In a book by Ruch that must be considered a classic in testing literature, entitled The Objective or New-Type Examination, 1/ there is a section, "Part III: Experimental and Theoretical Considerations", which reviews several studies concerning guessing as a test-taking behavior and also the total impact of the correction for guessing. A general survey of the research studies reported in this chapter, including some by Ruch (with others), seems to be that the correction for guessing of the standard type does, indeed, increase the reliability and the validity of the tests somewhat in most instances - although the gains are not great. However, no indication is given that corrected and uncorrected scores correlate 1.00 if all items are answered, as was subsequently realized.

This experimental section also reports on the effect of increasing the number of choices up to a maximum of seven and suggests that the five-choice item is probably the optimal number of choices to use where sensible alternatives can be found.

More recently, studies in the area of correction for guessing have been comprehensively reviewed and summarized in an article by James Diamond and William Evans published in the Review of Educational Research Spring 1973. 2/ In the summary of this article, immediately prior to the comprehensive listing of references, this sentence appears:

"By way of summary, one might note that the standard correction for guessing implies only one model of test-taking behavior. Perhaps new, computer oriented weighting procedures will allow us to expand the model and to consider other factors in test scoring, guessing, reliability and validity."

The bibliography included with this article under the various subtopics is both comprehensive as to time span and, in view of the length of the bibliography, indicates the continued concern with the problem of guessing in and up to the present moment - as suggested by the reference to new computer oriented procedures which will allow for weighting of test items to correct for guessing.

The study with which this report is concerned differs essentially from any of the studies reported so far in the literature - as nearly as can be discovered by a superficial review of the titles in the Diamond-Evans bibliography and some personal investigations of the author.

It has been customary to note in the most recent textbooks and other authoritative sources that the correction for guessing is totally ineffective when all items in the test are attempted, since the corrected scores and the uncorrected scores will have a correlation of 1.00.

For the benefit of those for whom this truth is not self-evident, Chart I-1 shows a bivariate of the actual scores of 62 cases attempting all items on one test versus the corrected scores for these same children. Note that the only variation in rank order is due to rounding off the corrected scores.

The article by Diamond and Evans in the review mentioned above indicates that under certain circumstances, other than the attempting of all items, the same phenomenon is true. In any case, the pursuit of a mathematical correction for guessing based on number right versus number wrong seems to be pretty much a lost cause; consequently, we must have some way of approaching the problem quite differently.

One obvious way would be to find a different approach to the identification of the child who actually guesses as compared to the person who has partial knowledge or has extensive knowledge and answers most of the questions from a basis of information exceeding that of his peers. To anticipate some developments which will be described later on, it is evident to the writer that the chances of finding any mathematical solution to this problem, at this time and under present circumstances, is quite unlikely.

The other long-term approach is to devise a way of testing which will be essentially free of guessing and, therefore, will cause the problem of guessing to disappear.

Obviously if one uses a work-sample method of testing, in which the child does the thing on which he is supposed to be being measured, then guessing is nullified - since he must perform the very task he is expected to perform in real life.

The best example is, perhaps, in arithmetic - where in a computation situation, such as one calling for the multiplication of two two-place numbers, the child actually does out the work and records his answer - possibly transferring the answer to a

1/ Ruch, Giles M. The Objective or New-Type Examination. Scott, Foresman and Co., 1929.

2/ Diamond, James, and William Evans. "The Correction for Guessing." Review of Educational Research, 1973, 43, 181.

Chart I-1

A Demonstration Bivariate to Show the One-to-One Relationship Between Corrected and Uncorrected Scores When All Items Are Attempted*

Actual Score

Corrected Score (Rights - One-third Wrongs)	-5 -2	-1 0	1-2	3-4	5-6	7-8	9-10	11-12	13-14	15-16	17-18	19-20	21-22	23-24	25-26	27-28	29-30	31-32	33-34	35-36	
35-36																					0
33-34																			1	2	3
31-32																					0
29-30																		1			1
27-28																	2				2
25-26																3	3				6
23-24																4					4
21-22															9						9
19-20														2							2
17-18													1	2							3
15-16																					0
13-14												5									5
11-12											1										1
9-10										1	2										3
7-8										2											2
5-6									5												5
3-4								4													4
1-2							3														3
-1 0							3														3
-5 -2					1	5															6
	0	0	0	0	1	5	6	4	5	3	3	5	1	4	9	7	5	1	1	2	62

r = .99

*Population: Random Sample - Word Meaning - Boys - Fall marginal answer space for ease in hand-scoring; but he does not come up with a response that is scorable by electronic methods.

The work-sample approach is, of course, older than standardized testing (as old as education, in fact). The first Stanford Achievement Battery, namely Form A published in 1923, used the work-sample approach in a number of instances, one of which was in the Spelling Test. In this test a paragraph was dictated to the child, all of which was recorded in writing by him. Only certain words in the paragraph were considered relevant in regard to their correct, or incorrect, spelling. Obviously, this was a time-consuming test to give and difficult to score - since the words had to be found in the context of the child's writing and then be evaluated as regards spelling corrections. In the next subsequent edition, this method of measuring spelling was dropped for a different approach.

While some ingenuity might increase the efficiency of the operation, it seems reasonably clear that the answer to getting rid of guessing is not to go over entirely to the work-sample approach but to do something to change the child's attitude toward testing in general and guessing in particular, while at the same time making it much more difficult to guess and still get a correct response.

A positive and logical approach to this problem is to analyze the types of content to be measured and then try to devise new item types which will satisfy the criteria mentioned above, while at the same time changing the attitude of teachers and pupils toward the administration and interpretation of the test results.

This proposal is a major task and this report is obviously not the place in which to discuss it in great detail. Suffice it to say now that the writer is convinced that the task is not an impossible one and before concluding this report he will attempt to indicate ways in which giant steps can be taken to effect this desirable goal.

The Purpose of the Study

Initially, the purpose of this study was to reveal by an intensive analysis of certain available data the extent to which guessing really existed and the nature of the groups who were most inclined to guess as a way of responding to the test situation. Any other findings were thought of as being more or less secondary.

Along with the identification of the children who guessed went the almost equally serious problem of constructing a test that could be given at the beginning of the year and at the end of the year with meaningful analysis of the differences between the two testing periods.

The author has been working exhaustively on this problem, which is really not so much statistical as it is logistical. The data from the study clearly shows that something must be done to replace the usual achievement battery for the purpose of a before-after type of testing over short periods.

Because of certain shortcomings, which will be developed more fully at a later time, the usual procedure for selecting standardized achievement test items won't work and an alternative procedure must be found.

The Available Data

In the 1969-70 school year, the State Department of Education in New Hampshire conducted a statewide testing program involving the Stanford Achievement Test. The inclusive testing program covered grades 2, 4, 6 and 8, but this report is concerned only with grade 4 - in which Intermediate I Battery: Form X was used.

It was further stipulated that the same form of the test given in October to all pupils in the state would be given over again in the spring to pupils identified as being in Title I projects. The spring testing of these Title I children was supposed to provide the data in terms of which to evaluate the effectiveness of the instruction in the Title I projects as compared to the normal amount of growth during this period of time.

Typically, this "normal amount of growth" is expressed in terms of month of grade equivalent and it has been considered satisfactory to set up some criterion, such as growth of one school year during the seven-month period, as being the expectancy in a successful program.

Without getting into all of the complexities involved, grade equivalents are totally unsuitable for this purpose and always have been. They are based upon testing over a period of twelve months even though the amount of gain from one testing period at a given grade, such as grade 4, to the next testing period, at grade 5, is twelve months. The fact of summer forgetting is neglected totally, and it is assumed that a month of grade increment for reading means the same as an increment of one month in arithmetic, i.e. the rate of growth from subject to subject is constant. All of these assumptions have been proved to be totally incorrect.

About the time this program was getting under way, this writer was asked to act as a consultant to the State Department of Education and, specifically, to Title I within the State of New Hampshire to help implement a program that would be effective. The first step that was taken along these lines was to persuade the Department of Education to re-test in the spring a random sample of pupils from the entire state as a control group so that the gains made by the Title I pupils in the state could be compared with gains made by this random, and thus representative, sample for the state.

This was an enormous step forward, especially in this context when there were no spring norms and fall norms available for the Stanford Battery. Stanford was typically standardized in the spring, as Metropolitan (1/) is typically standardized in the fall, and extrapolations over the period of time from school year to school year did not provide a satisfactory method of determining the amount of growth to be expected over a seven-month period - especially in the case of a test having a variety of subtests.

The identity of the children who were to comprise the random sample was determined by use of a random technique employing the IBM-360 Model 50 computer at the University of New Hampshire. The work was done under the direction of the Bureau of Educational Research and Testing Services. 2/

It was specified that a total of 1,000 children out of about 10,000+ were to be identified to constitute this random sample. These children were further identified by school and a request was sent to the administrators of the school districts in which these children resided to have them re-tested at the same time the Title I children were re-tested in the spring.

Apparently not all of the schools chose to comply to this request, so that the total number of children actually tested for the random sample was a little more than 600 - as compared to the specified 1,000. The experimental population, if one wants to call it that, consisted of 426 children in the Title I program, concerning which much more will be said later in a separate section.

Evaluating the Random Sample

The scores for the fall testing program made by the children actually drawn for the random sample for whom results were available both fall and spring were distributed on a number of variables and their results were compared with the total state sample. The outcome of these comparisons is given in detail in the report to the State Department of Education entitled "A Description and Evaluation of the Statewide Testing Program in New Hampshire in 1968-69 and 1969-70 under the Sponsorship of Title I and the Significance of the Data Obtained for Evaluation with this Activity." This report was completed in July 1971 under a contract of March 13, 1971.

It would be too space-consuming and redundant to reproduce in its entirety the comparison of the random sample available to us for analysis with the total group for the state, but it can be said that every type of evidence of comparability obtained seemed quite convincing that the two populations were sufficiently interchangeable for our purposes.

In Chart I-2 the random sample IQ distribution is reproduced on a Normal Percentile Chart superimposed on the IQ distribution for the state as a whole. It is obvious by looking at the two distributions that they are very nearly the same. Statistical tests might have been suitable in this situation but were not thought to be necessary because all that was really needed was a population tested both fall and spring that was reasonably representative of the state as a whole.

Random Issues Relative to Testing

Even if Stanford Form X were a criterion reference test (which it was not) in which it was expected that most or nearly all of the children would answer most of the questions - or nearly all of them - right, even this then would leave unanswered the extent to which the ablest children still are not learning as much as they are capable of doing or that they are learning the right kinds of things for them considering that they are atypical with respect to the grade as a whole. The same generalization is

1/ Metropolitan '70 has both fall and spring standardization programs.

2/ At this point, the writer would like to express his appreciation to Mr. Richard Clukay for his great help in preparing programs, debugging them, and implementing the analysis of the data on the computer - as indicated in the Title I Report as mentioned above.

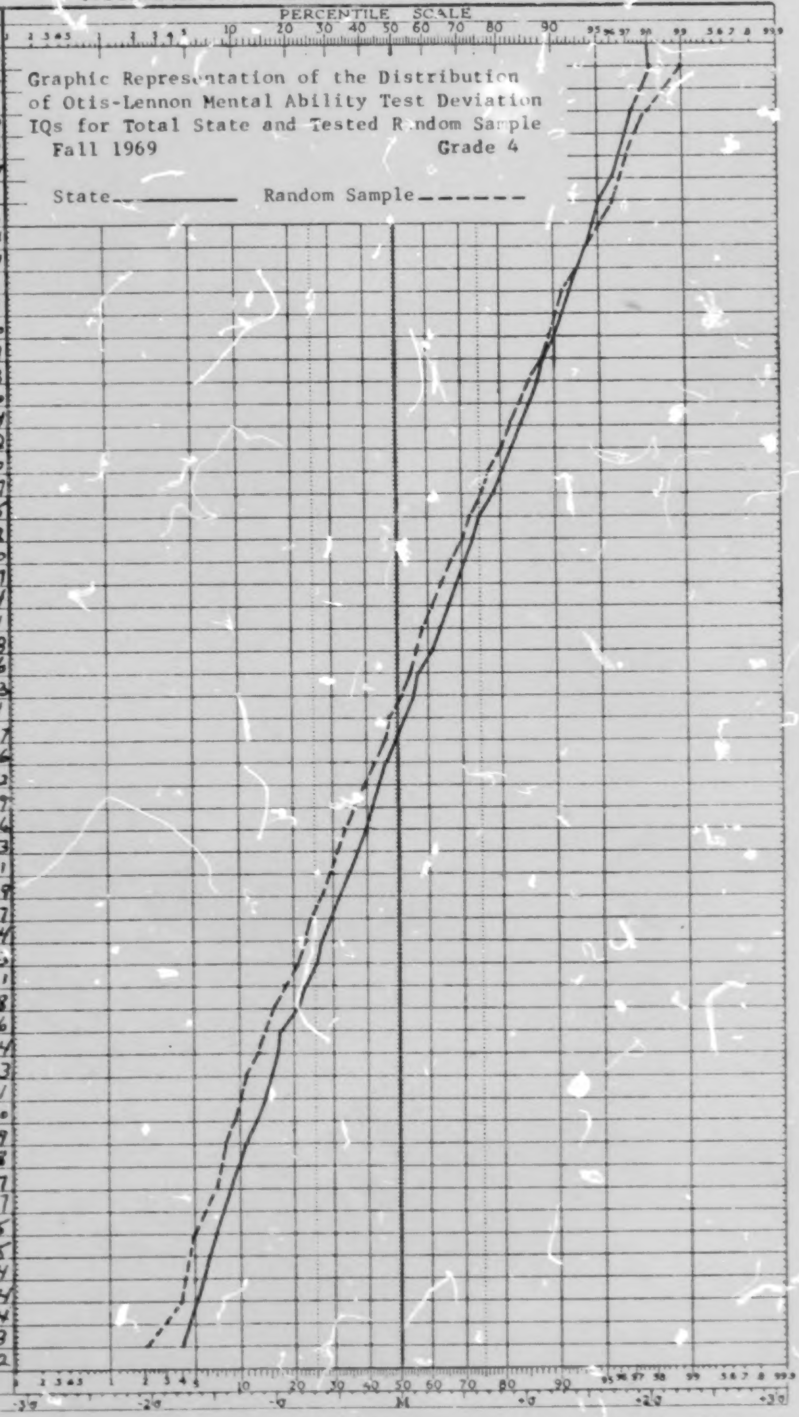
Answer Sheet Study - I

Chart I-2

NORMAL PERCENTILE CHART

By Arthur S. Otis

Otis-Lennon IQ	State-wide 11,938			Random Sample 585		
	I			II		
Score intervals	Frequencies	Subtotals	Per cents	Frequencies	Subtotals	Per cents
131+	257		98	7		99
130	53		98	2		99
129	45		97	4		98
128	70		97	3		98
127	42		96	2		97
126	65		96	3		97
125	53		96	5		96
124	100		95	7		96
123	87		94	4		94
122	113		94	6		94
121	117		93	8		93
120	100		92	5		91
119	139		91	7		90
118	144		90	7		89
117	163		88	12		88
116	168		87	10		86
115	186		86	15		84
114	216		84	11		82
113	319		82	17		80
112	245		80	9		77
111	306		78	17		75
110	248		75	17		72
109	[illegible]		73	13		70
108	269		71	17		67
107	357		69	18		64
106	279		66	18		61
105	284		64	12		58
104	454		61	20		56
103	286		57	10		53
102	391		55	23		51
101	240		52	8		47
100	339		50	24		46
99	342		47	18		42
98	270		44	16		39
97	256		42	15		36
96	315		40	13		33
95	248		37	11		31
94	215		35	13		29
93	317		32	16		27
92	257		30	8		24
91	222		27	9		23
90	277		26	17		21
89	286		23	15		18
88	291		21	10		16
87	219		18	9		14
86	210		17	11		13
85	144		15	6		11
84	211		14	7		10
83	143		12	6		9
82	129		11	2		8
81	109		10	2		7
80	118		9	8		7
79	124		8	3		5
78	79		7	3		5
77	72		6	2		4
76	95		5	3		4
75	63		5	4		4
74	58		4	4		3
-73	421		4	13		2
Medians	101			102		



Graphic Representation of the Distribution of Otis-Lennon Mental Ability Test Deviation IQs for Total State and Tested Random Sample
Fall 1969 Grade 4

State —————— Random Sample - - - - - -

Published by World Book Company, Yonkers-on-Hudson, New York, and Chicago, Illinois. Copyright 1938 by World Book Company. Copyright in Great Britain. All rights reserved.

certainly true of the Title I children who were, by selection and definition, also an atypical group.

It is nice to know that they (and most of the other children) are learning or have learned, at one time or another, everything considered to be relevant for inclusion in a criterion reference test, but it does not tell the teacher about what to do next for the exceptional child. How much farther could the exceptional child go on from there if his opportunities were less limited as he lock steps his way through a prescribed minimum foundations curriculum? Contrariwise, do Title I children master the learning essential for them to master before going on to new material?

Sometimes it is possible to use a well-standardized general achievement examination as the basic core for a testing program which is supplemented by locally-made test items to fill in the gaps covering learnings thought to be essential by those concerned with curriculum matters at the local level. Thus, the achievement test can be interpreted in terms of the norms provided for it, and an item analysis can be done to give information concerning performance on individual items.

To the score on the published test is added a score on some supplementary test intended to round out the inadequacies of the standardized survey-type instrument, and the total score plus item analysis of all items is taken into account in determining the adequacy of the instruction in light of the needs of individual pupils within the school, class, or instructional group.

The most significant use of criterion reference testing should be found where teaching most desperately needs to be individualized; namely, with those children who show a disparity in their performance from what is typical of their peer group. In these instances, it is then possible to go on with a higher level of instruction for some or to slow up the pace until mastery reaches the level established for others.

Any idea, however, that children are universally going to master the material typically found in a textbook or recommended for teaching at a particular grade is just not going to happen. Sometimes it takes as much as two or three grades beyond the grade level at which a topic is introduced before it is really adequately learned so that it becomes a part of the tool kit for the child in attempting to learn or to attack problems at a higher curriculum level.

Many, probably most, communities or school systems take their responsibility for instructing teachers in the intricacies of test item writing very lightly, feeling that this is a task that should have been accomplished during the course of their undergraduate training.

The fact of the matter is that training in the development of tests is one of the most neglected areas in teacher education, and few new teachers enter the classroom prepared to undertake the simplest kind of test construction that could be considered to be scientifically valid - to say nothing of interpreting the results of tests (especially ones supplemented by local items) in a manner that is consistent with the best statistical and methodological practices in educational and psychological measurement.

Advocates of mastery teaching, which is the natural outgrowth of the criterion reference approach to educational evaluation by test, recommend keeping a child working on a particular knowledge or skill, or coming back to it very frequently, until he can demonstrate what they consider to be a satisfactory mastery of that particular item or that skill. The writer has considerable sympathy with this point of view if, and when, it is possible to establish an hierarchy and to demonstrate that it is essential that persons know certain material at a particular grade or development level before they should go on to another still higher level of instruction.

In the 1890's it was common for schools to use textbooks which were graded, not in the sense of being assigned to a particular grade (as this term is used in this country now) but were sequential, and a child was required to stay in a particular "book" until he had mastered the content of that book before he was allowed to proceed to the next one. In those days schools were small, often being of one or two rooms, and the teacher could handle this type of situation because the older children became the teachers of the younger and many children learned in school by listening to their older brothers and sisters recite as called upon by the teacher.

With the modernization of education and the development of a grade system, this practice was, of course, abandoned. Now we find ourselves coming back to a kind of structuring of the curriculum and of instruction that closely corresponds to the olden days or, in more modern terms, closely corresponds to programmed instruction (or the procedure used in programmed instruction using its test, teach, test psychology). In programmed instruction, a child is rarely supposed to proceed to the next higher level until he has successfully answered test items (supposedly indicating mastery of the knowledge or skill currently being taught) and the hierarchy, whether or not it does in fact exist in truth, is there because it is imposed by the person who is responsible for constructing the programmed text.

It seems to make a great deal of sense, however, for us to re-examine the entire curriculum and, insofar as possible, to break it down into behavioral objectives or performance objectives which can be shown to follow some hierarchy (even an artifical one). Knowledges and skills which are peripheral as such should be treated as such, allowing the child in a class to learn everything he can learn about the world in which he lives - whether it is formally considered a part of the curriculum for which the child should be responsible at that point in his development, or whether it simply is a way of broadening his understanding of his world.

Individualization of the curriculum in the American public schools is probably the trend of the future, but if this is the case one must face up to the fact that it may mean enormous increase in costs of public education, either for instructional personnel or for equipment which will substitute for instructional personnel - such as audio-visual aids, computers and the like.

Administrative changes which allow pupils to move through the curriculum at their normal pace can contribute to the individualization of education, but in the long run someone must make the decision as to what the imperatives are and to see to it that they are provided for.

Thus, if an instrument was constructed for the purpose, one could measure the extent of the gain subtest by subtest over a short period. Comparisons could be made for any other subgroup breakdown, such as boys versus girls, over the seven-month period separately for each subtest, and one could evaluate this gain in raw score points or in standard score points, but certainly not in terms of grade equivalents or percentile rank.

Organization of the Study

The organization of this report is such that the results of the administration of Stanford Intermediate I Battery: Form X in the fall and spring for the random sample will be presented first in several ways, allowing the reader to draw whatever conclusions he wishes concerning present success or failure from these comparisons. Beyond our own commentary, we will present alternatives as to the configuration of any battery to be used in such a manner.

Methods of Data Analysis

At this point, it is quite essential to call attention to some significant and unusual aspects of the data finally available for analysis.

It is not the common practice to repeat the same form of a test over a period of time to measure gain because of the possible effects of remembering the answer given on the first administration. Factors of finance and logistics, I think, were uppermost in the minds of the State Department of Education when the decision to use the same test was made. It was decided to leave the test battery in the hands of the local school administrators so that only answer sheets would need to be distributed in the spring - and thus the distribution problem and the matter of determining the real equivalence of two forms purported to be comparable would be bypassed.

Regardless of the general merits of this procedure, with which the writer was first inclined to totally disagree, it did afford an opportunity to make a type of analysis that could not otherwise have been made; namely, the comparison of the response made by each child to each item on each test fall and spring, so that it was possible to study the consistency of response from fall to spring in a very detailed manner.

However, before presenting these data it would be well to take a look at the amount of the gains that were found in terms of raw scores for the fall as against the spring tests separately by subtest. For this purpose it was, therefore, only necessary to compare the random sample fall testing results with the spring results for the same group of pupils. This comparison, in other words, did not involve Title I children because what was being attempted at this point was merely to determine what was generally a normal or typical gain so as to provide a basis for evaluating subsequently what Title I children did under the same conditions.

In order to accomplish what we need to know for this report, we only have to reproduce a portion of the data appearing in the original statewide report, previously referred to, which provided for the comparison of raw scores for fall and spring together with the raw score gains. This table also gave the amount of gain in terms of grade equivalents - to satisfy the "believers" in this approach, especially the U.S. Office of Education!

Since this study is restricted to grade 4, the data reported herein are for grade 4 only and for the five selected subtests - which were all it was felt we could handle in the present project. Our Table I-1 is, therefore, reproduced from Table VII-B-1 of the previous report.[1] It is essential to remember that the gains reported for the random sample are over a seven-month period only. These gains are reported separately for the 75th, 50th, and 25th percentile ranks, but the main emphasis is on the median, or 50th, percentile rank.

An examination of the raw score gains, considering the medians only for the moment, shows that they amount to perhaps a point per month of instructional time for Word Meaning, Paragraph Meaning, and Arithmetic Computation, but drop substantially to four raw score points in Arithmetic Concepts and Applications. This drop is due, in substantial measure, to the smaller number of items in the last two arithmetic tests.

In terms of grade equivalents, the gains also appear to be about as one would expect them to be, except that the gains in Word Meaning and Paragraph Meaning exceed what one would expect over a seven-month period, i.e. a gain of about one point per month of school instruction.

Perhaps the most distressing fact coming out of this comparison is the relatively small difference between children at the 75th percentile rank versus the 50th. Even the difference between the 25th percentile rank and the 75th is small. In Word Meaning, for example, there is no raw score gain, i.e. both of these percentile ranks have comparable percentiles, or raw scores, of 6. In Paragraph Meaning the difference is three points, from 7 to 10, with a difference of only one point between the 25th and 50th percentile in a sixty item test.

Perhaps this was our first clue to the fact that these tests really did not do a very good job of measuring, since the spread of scores for the middle 50% of children in the group was not very large.

The evidence just discussed filled this writer with apprehension as to what would be found when Title I children were analyzed in a similar fashion. This led to some further investigation of results for the random sample before proceeding with any investigation of Title I.

New Data Available for Analysis

At this point, we must shift our attention to the additional data wrung out of the testing program by use of the answer sheets, which were still available at Measurement Research Center for both fall and spring testing. These answer sheets were obtained and fall and spring sheets were matched up, using available code numbers, with the result that we came up with the numbers of cases previously mentioned as a basis for comparison. These are cases for whom we do have complete data for both those included in the random sample and for Title I children.

Just what the loss of cases due to incompleteness of individual pupil data did to the analysis is, of course, impossible to tell, but we have some data bearing on the tested random sample versus the population analyzed which seem to establish rather firmly the fact that the sample drawn randomly and actually tested gives a remarkably close representation of the performance of the state as a whole.

With the data on IBM tape available for computer analysis, a whole new list of analytical approaches was open to us - many of which have been completed and are reported in the subsequent pages of this study, but many more "in the wings" if further analysis becomes possible.

The rosters on which the item analysis data for this study were reported gave the fall and spring test response for each of the items in each test included in the study. These data were listed by the actual response number marked by the child - not in terms of rights, wrongs, and omits.

A key was included as an extra line on the listings after every fifth pupil's fall and spring item analysis data. The computer printout constituting Figure I-1 shows a typical page from the random sample listings. Similar pages were available for the Title I listings.

[1] "A Description and Evaluation of the Statewide Testing Program in New Hampshire in 1968-69 and 1969-70 Under the Sponsorship of Title I and the Significance of the Data Obtained for Evaluation With This Activity." Prepared by the Test Service and Advisement Center. 1971.

Answer Sheet Study - I

The page from the random sample was selected to be as representative as possible for the whole sample and it reveals some surprising information concerning test-taking behavior, especially as regards guessing tendencies. Some of it is so subtle that it defies analysis in terms of the usual types of statistical summarization but depends, indeed, on a critical visual examination of the data, i.e. what is sometimes called "eyeballing." This relates in particular to the pattern of scores that subtly differentiates the guessing child from the child who rarely guesses. This is the changing pattern of marking from a succession of "rights" at the beginning to a pattern typical of a "guesser," or a chance pattern.

Table I-1

Percentiles and Grade Equivalents
Corresponding to Selected Percentile Ranks

RANDOM SAMPLE

Test	Items	%ile Rank	Raw Scores Fall	Spring	Gain	Grade Equivalents Fall	Spring	Gain	Dev.*
Word Meaning	38	75	21	27	6	4.9	5.9	1.0	+.3
		50	15	22	7	3.9	5.1	1.2	+.5
		25	10	16	6	3.3	4.1	.8	+.1
Paragraph Mean.	60	75	30	40	10	4.6	5.9	1.3	+.6
		50	23	31	8	3.8	4.7	.9	+.2
		25	17	24	7	3.0	3.9	.9	+.2
Arith. Comp.	39	75	14	23	9	4.0	5.2	1.2	+.5
		50	11	18	7	3.6	4.5	.9	+.2
		25	8	13	5	3.1	3.8	.7	0
Arith. Concepts	32	75	16	20	4	4.8	5.5	.7	0
		50	12	16	4	4.1	4.8	.7	0
		25	9	11	2	3.3	3.9	.6	-.1
Arith. Applic.	33	75	16	21	5	4.6	5.5	.9	+.2
		50	12	16	4	4.0	4.6	.6	-.1
		25	9	11	2	3.6	3.9	.3	-.4

*Represents the Deviation from the Expected Gain of .7 of a Calendar Year, often inaccurately designated as 7 months of a School Year.

FALL AND SPRING ITEM ANALYSIS CHART PAGE 004
STANFORD ACHIEVEMENT TEST - INTERMEDIATE I - GRADE 4 - WORD MEANING
NEW HAMPSHIRE STATEWIDE TESTING PROGRAM - 1969-1970 - RANDOM SAMPLE MALES

ITEM NUMBER		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	TOTALS		
KEY		4	3	4	3	2	1	2	2	3	2	2	3	1	4	1	1	4	1	3	1	1	2	3	4	3	2	4	4	2	3	2	4	1	4	4	3	1	2	1	X	0
PUPIL	F-S																																									
097	F	4	3	4	1	2	2	2	2	3	2	3	2	1	4	3	1	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	12	5	21
097	S	4	3	4	3	2	2	2	2	3	2	2	2	1	4	1	1	4	1	2	1	2	2	2	4	3	2	4	3	2	2	2	1	1	3	2	1	4	2	26	12	0
100	F	4	2	4	3	2	3	2	3	3	2	2	1	1	4	4	1	4	4	4	1	1	2	4	1	3	0	0	0	0	0	0	0	0	0	0	0	0	0	16	9	13
100	S	4	3	4	3	2	1	2	3	3	2	1	4	1	3	4	3	4	1	3	1	1	2	2	1	3	2	4	3	2	3	2	3	1	4	3	2	3	4	24	14	0
101	F	4	3	2	1	4	3	3	1	4	4	3	3	2	3	4	3	2	1	2	4	1	3	4	1	4	3	3	1	4	2	1	1	2	4	1	1	4	1	6	32	0
101	S	4	3	4	2	2	3	3	1	3	1	1	4	1	3	4	3	4	1	2	2	1	2	3	1	3	4	4	4	2	3	1	1	2	4	3	1	2	4	17	21	0
102	F	4	3	4	3	2	1	2	2	3	2	2	2	1	4	1	1	4	2	2	1	4	3	3	4	2	0	0	0	0	0	0	0	0	0	0	0	0	0	19	6	13
102	S	4	3	4	3	2	1	2	2	3	2	2	3	1	4	1	1	4	1	2	1	2	3	3	4	3	2	1	1	4	3	2	4	1	4	2	4	1	4	29	9	0
106	F	4	3	3	3	2	2	2	1	3	2	3	3	1	4	4	1	4	2	3	4	2	2	1	4	3	2	4	0	0	0	0	0	0	0	0	0	0	0	18	9	11
106	S	4	3	4	3	2	1	2	2	3	1	2	2	1	4	1	1	4	1	2	4	1	3	1	2	3	2	2	3	4	1	1	2	1	2	2	1	2	4	20	18	0
** KEY **		4	3	4	3	2	1	2	2	3	2	2	3	1	4	1	1	4	1	3	1	1	2	3	4	3	2	4	4	2	3	2	4	1	4	4	3	1	2			
107	F	4	3	4	1	2	1	2	2	3	2	2	3	1	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	13	1	24
107	S	4	3	3	3	2	1	2	2	3	2	2	3	1	4	1	1	4	2	2	4	2	3	2	1	1	2	1	4	4	0	0	0	0	0	0	0	0	0	18	11	9
108	F	4	3	4	3	2	2	2	2	3	2	2	3	1	1	1	1	4	1	3	1	2	3	2	4	3	2	4	4	2	3	1	4	1	2	1	1	3	3	27	11	0
108	S	4	3	4	3	2	1	2	2	3	2	2	3	1	4	1	1	4	1	3	1	4	1	3	4	3	2	4	4	2	3	2	4	1	2	4	1	3	4	32	6	0
109	F	4	3	0	3	1	2	4	2	0	2	4	1	2	4	2	1	2	3	4	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8	11	19
109	S	4	3	4	4	4	1	2	2	4	2	1	3	3	3	4	3	4	1	2	4	1	3	3	3	4	2	1	2	2	3	2	2	3	4	4	4	1	3	19	19	0
113	F	4	3	4	3	2	2	2	1	3	3	2	3	1	3	2	1	4	1	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	14	5	19
113	S	4	3	4	3	2	2	2	1	3	2	2	1	1	3	1	2	3	1	3	1	3	3	3	4	3	2	2	4	2	2	0	0	0	0	0	0	0	0	20	10	8
114	F	4	3	4	3	3	1	2	2	3	2	1	3	1	1	1	3	4	1	3	1	1	2	2	2	3	1	0	0	0	0	0	0	0	0	0	0	0	0	19	7	12
114	S	4	3	4	3	4	1	2	2	3	2	2	3	1	3	1	1	4	1	3	2	4	2	3	3	2	2	1	1	2	1	2	0	0	0	0	0	0	0	22	9	7
** KEY **		4	3	4	3	2	1	2	2	3	2	2	3	1	4	1	1	4	1	3	1	1	2	3	4	3	2	4	4	2	3	2	4	1	4	4	3	1	2			
117	F	4	3	4	3	2	1	2	2	3	2	2	1	1	3	1	1	4	0	1	1	0	2	3	0	1	4	0	0	0	0	0	0	0	0	0	0	0	0	18	5	15
117	S	4	3	4	3	2	1	2	2	3	2	2	4	1	4	1	1	4	1	3	1	1	3	3	4	3	2	4	3	2	3	2	1	1	3	1	1	1	0	30	7	1
118	F	4	3	4	3	2	1	2	2	3	2	2	3	1	4	1	1	4	1	3	1	1	2	3	4	3	2	1	0	0	0	0	0	0	0	0	0	0	0	26	1	11
118	S	4	3	4	3	2	1	2	2	3	2	2	3	1	4	1	1	4	2	3	1	1	2	3	4	4	2	4	4	2	3	2	3	1	4	4	3	1	2	35	3	0
119	F	4	3	4	3	2	1	2	1	3	2	4	1	1	3	4	3	4	2	2	0	1	2	1	4	3	2	3	4	2	0	0	0	0	0	0	0	0	0	18	10	10
119	S	4	3	4	3	2	1	2	2	3	2	2	2	1	4	1	1	4	1	3	1	4	2	3	4	3	2	1	1	2	3	2	1	3	2	2	1	1	4	28	10	0
123	F	4	3	4	3	2	1	2	2	3	2	2	3	1	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	14	0	24
123	S	4	3	4	3	2	1	2	2	3	2	3	3	1	4	1	1	4	1	3	1	1	2	3	4	3	2	3	4	2	3	2	3	1	4	0	0	0	0	31	3	4
126	F	4	3	4	3	2	1	2	2	2	2	2	1	1	4	4	3	4	3	4	4	2	1	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	14	9	15
126	S	4	3	4	3	2	1	3	2	3	2	2	3	1	1	4	1	4	1	3	1	2	2	3	1	3	2	4	2	0	0	0	0	0	0	0	0	0	0	22	6	10

1=CORRECT,X=WRONG,O=OMIT

Test Scores Related to Information Theory

In information theory, there is a "sender" of a message, a "receiver" of this message, and other interfering "noises," or "static," which keep the message from being clearly or completely understood. The extent to which this "static" is present is a measure of the extent the communication is impeded; or, if it is absent, is a suggestion that the message communication is entirely perfect (a rare phenomenon).

When applied to tests, information theory postulates that the basic purpose of a test is to provide the medium by which a child can communicate to his teacher (or to others) what he truly knows and what he does not know. In this situation it would seem to be evident that anything that even remotely smacks of guessing must be considered in the nature of "static" - because it clouds the validity of the message the teacher is receiving from the pupil. In a nutshell, this constitutes the most important phase of this study when it is taken overall; namely, the identification of ways in which the "message" may be conveyed from the pupil to the teacher with the least static.

Obviously this may involve great changes, not only in the nature of the tests being used but also in the conditions under which they are administered. It certainly calls for a climate of confidence within the classroom which will allow the child to feel free to respond to an item or not to respond to it; or hopefully, in the latter case, by use of the unambiguous "Don't Know" space - which should be provided to permit him to indicate to the teacher that he specifically does not know the answer to the question being asked.

A summarization of this "Don't Know" information may, indeed, be the most important data coming to the teacher as a result of the test item analysis.

What this writer has to say subsequently concerning the analysis of these data is clearly and purposefully reflective of this basic point of view; namely, that testing does constitute in the educational field a kind of communication. The significance of our success or failure is that the final validity of the test rests on our ability to do this. Even more important is the confidence the teacher can have that the student's response to a particular item is truly indicative of his grasp of the material being tested, both for the individual and in the needed groupings of this information. The child's true position relative to his peer group is also at stake.

The need for such normative interpretation seems perfectly obvious. However, before we get to the point of interpreting our scores in terms of norms we should emphasize once more that the building block of a test is the test item; that the test item comes directly out of the context of the generally available instructional material; that the test-maker does not choose what items he shall retain or leave out of the experimental or tryout test but rather chooses for the final test items which are known by enough children to make it worthwhile to include them and eliminates some items which are wholely mastered - probably because they are below grade level for the time of year and the grade level at which the test is used.

In this study, the Stanford Intermediate I Test apparently gave very good results at grade 4, in terms of conventional measurement criteria, considering the fact that the subtests are unduly short - having less than forty items in every test except Paragraph Meaning. The distributions of scores are fairly symmetrical and show the characteristics that a measurement person normally looks for when he is attempting to make use of a group test of this sort for measurement purposes, i.e. reasonable statistical reliability and a generally satisfactory "status report" on the ablest and least able pupils as judged independently of the teacher's evaluation. 1/

For curriculum purposes, however, the test was not long enough for the ablest pupils and range of difficulty too steep for the least able - as will come out in the course of our subsequent investigations.

More About Criterion Reference Testing

One of the really active movements in the field of testing in recent years has been the development of what is known as the "criterion reference test," which presumably is a test which reveals what the child has learned at the particular grade level at which he is functioning or earlier. The theory behind this approach is that the criterion reference test, in contrast to the general achievement test, will reflect skills and knowledges of such paramount value that everyone, or nearly everyone, in the group should be expected to answer the questions correctly, i.e. to demonstrate mastery of these stepping-stone atoms of in-school instruction.

1/ SAT national norms turned out to be on the hard side for no discernible reason, but the preferred use of local norms sidestepped this shortcoming.

This is not the time and place to discuss criterion reference testing in detail, except to emphasize the fact that the approach taken in this study is somewhat similar to the approach that would be taken in a criterion reference test in that more attention has been paid to the performance of individual pupils on single items than is normally the case in analyzing the results for a general achievement battery. Thus, the question of item validity is a paramount issue.

Many people have raised the question as to whether general achievement batteries may be used in the place of criterion reference tests. This is a difficult question to answer because it depends upon the extent to which the general achievement test used is properly graded for the group taking it. Some tests, for example, may be too easy for the least able or even typical fourth-grader and definitely too easy for an exceptionally, i.e. above average, able group - not providing enough difficult material to reveal the top ability level of the "most able" children.

Who should see that these imperatives are appropriately taken care of and not lost in the welter of activities which makes up the ongoing daily program of the public schools is a real question. For example, a child can hardly be expected to cope with the curriculum of the middle school or the junior high school who has a reading level of only average grade 3 pupils.

Who is to monitor a pupil's progress and on the basis of what objective data?

There are other implications growing out of the author's contention that the test situation is essentially one of communication between the child and the teacher and vice versa. If this is true, tests and test results should play an important role. The most obvious implication of this philosophy is that the child should, by all means, know whether he answered "the question" correctly or whether he missed it and, if he did miss, that he shall be given an opportunity to review or have additional instruction, as needed, in order to learn the knowledge or skill and get recognition for so doing - unless the item in question is dismissed as one that is peripheral to the course curriculum or the imperatives of the curriculum.

Similarly, it implies that the teacher must take the trouble of finding out who knows what, i.e. group analysis, and to provide a program for those children who have mastered a particular item or skill or a group of items or related skills, and a differentiated one for those who are in need of additional learning material as well as the encouragement for others to go on to learn at the rate commensurate with their abilities.

Teachers will say that to take such supplementary action for children who are atypical involves an effort on her part that is unreasonable to expect in light of today's demands on people outside the realm of their way of making a living. Remember that about two-thirds of the children within a particular classroom that is heterogeneously grouped will be quite similar in their learning potential. There is no answer to this objection on the part of the teacher except to provide the help needed, as indicated above, in a suitable manner. In some cases, there certainly is no way to avoid the necessity of the teacher aide or some human being to be there to meet a crisis or provide a learning situation as needed.

There is no doubt that the ideal situation in teaching is, as is so often said, "Mark Hopkins on one end of the log and the pupil on the other." Such a situation necessarily implies the desired debate or sharing of knowledge between pupil and teacher so that not only is a fact learned or a skill mastered but the child knows why the given fact is true and why other possible answers are not true if there is a choice.

The able child in such a learning situation is then encouraged to move ahead as rapidly as he can, while the slower child is dealt with patiently and is given the requisite practice and drill in learning those prerequisite knowledges and skills while not being deprived of his share in the fun that is an essential part of going to school.

Treating test data as if it constituted the transmission of a message from pupil to teacher and vice versa has implications with regard to the climate of confidence in the classroom. It must lead, inevitably, to a thoroughgoing consideration of restructuring American education in the manner indicated above, to provide for these multiple levels of accomplishment at each major mileage marker which designates the place where the child should be in terms of his mastery of the established hierarchy of knowledges and skills up to his level of learning ability.

Differences in effective learning rate will not go away regardless of fervent arguments that they are environmental, not inherited. The fact that must be dealt with is that they are there - real, measurable, and constantly influencing learning.

Actually, the absolutely basic knowledges and skills in the lower grades center around the development of reading skills, the development of vocabulary, and the development of ability to compute to the point where number combinations and the like are automatic. It is not the purpose of this paper to develop a scheme for accomplishing all this in an administrative sense, but simply to provide evidence arising from the present testing program, described above, that this is not being done if the way children answer test questions may be taken as evidence of this failure of their knowledge.

Certainly, this study implies that guessing in a multiple choice item situation has no place in testing and everything possible must be done to insure that the child is encouraged to give an honest response which, more often than not for some, may be a "Don't Know."[1]

The problem of time limits in test-taking also is a situation that must be dealt with more imaginatively. Generally the practice is, in the construction of published achievement tests, to arrive at a time limit such that all but one or two pupils in the class will have an opportunity to do all that he can do in the time allowed. This assumes that the test items are arranged in order of difficulty and that this order is more or less stable from one population to another, which is generally but not always the case.

Ways are suggested in this study of compensating for less than adequate time limits by estimating a total score on a time limited test using essential item analysis information available, according to the procedures to be recommended, and ranking the child as to his performance in school on the basis of this estimate rather than his actual score. Measures of school learning ability similarly should be developed which do not depend solely upon how many items a child can mark, correctly or not, within a given length of time. The prediction procedure to be suggested in the following pages is quite as applicable to such learning ability measures as they are to achievement tests.

---

[1] In the 1958 edition of the Metropolitan Achievement Test, the writer introduced for the first time (to his knowledge) the "Don't Know" space as an option. It was not to be scored but used as a way of keeping a child's test response intellectually honest.

It is best illustrated by its application in the Arithmetic Computation Test. In the handscoring edition of this test, the work was done in the booklet (work sample type) and the child transferred his answer to the margin of the sheet - where it was scored with a strip key but with the teacher looking back at the work done by the child in the process of scoring or subsequently.

The Spelling and Language Tests in this battery, for which the author was also responsible, made similar use of the "Don't Know" space. The basic principle involved was the same; namely, to provide a way for the disadvantaged or unknowledgeable child to escape the trap of having to answer randomly by marking the "DK" space as preferable to sitting for a substantial period of time doing nothing.

In the machine scoring edition of the test, the Directions for the Computation Test specify that the child shall actually do out the work, as before. The publishers offered to the user an "Arithmetic Worksheet" or optionally suggested that scratch paper could be used. The Directions for Administering: Arithmetic Computation (Machine Scored edition) say:

> "Work each example on the paper provided. As soon as you have worked an example, find the three answers given for the example in the right hand column of the test booklet. Then, on the separate answer sheet, fill in the space under the letter of the answer which agrees with yours. If you do not find your answer in the test booklet, fill in the space under NG (for not given) on the answer sheet. If you do not know how to work the example, fill in the space under DK (for don't know)."

The child's actual computation was to be left with the teacher with his work intact, while the answer sheet was sent for machine processing.

Note that in the machine scored edition three possible answers are given and, in addition to these three answers, an "NG" (Not Given) response is provided as well as the "DK" or "Don't Know" response. The NG response was a scored response, but the number of items so keyed in the test was minimal.

In this writer's opinion, the scratch paper was not a viable alternative because of the time required to copy the computation problem; but expediency won out.

If one analyzes this procedure closely, it is seen that in essence this is a job or work sample, and the marking of the separate answer sheet is merely a clerical task transferred to the child in addition to the work he has to do in making his computation.

The 1970 edition maintains to some extent the characteristics of the 1958 edition, but the separate consumable Arithmetic Worksheet is not available.

The joker is that both the 1958 and '70 editions were standardized using expendable booklets. Children were permitted to do the work in the booklet without having to copy off the examples, and the norms for Metropolitan were based upon this assumption.

SOME PERTINENT FACTS IN REGARD TO THE STANFORD ACHIEVEMENT TEST: INTERMEDIATE I BATTERY: FORM X

It is impossible, in the short amount of space available in this report, to cover all of the essential information concerning the Stanford Achievement Test, 1964 revision. Few people who are acquainted with achievement testing, especially of the battery type, can be found who are not familiar with the Stanford Achievement Test Series in general. It was the very first such general achievement test published and its publication date of approximately 1923 put it well ahead of its competition in regard to such battery-type tests.

The next major revision was in 1940, followed by another in 1953 and, finally, by another in 1964. At the time this study was undertaken, the 1964 battery was the current battery in use. It has subsequently been revised and the new forms became available in the fall of 1973.

However, the little attention that we can pay to the characteristics of the battery in this article must be confined to the 1964 edition, Intermediate I: Form X, and to the tests in Word Meaning, Paragraph Meaning, Arithmetic Computation, Arithmetic Concepts and Arithmetic Applications only.

This means that no data are reported here concerning the Spelling Test, the Language Test, the Social Studies Test, or the Science Test - although data of a similar nature to that basically used in this study are available for these other tests in the fall. Considerations of time and expense precluded the use of all of the tests in the spring and, of all of the tests in the battery, the ones that seemed to be most relevant and of most interest to the user were the tests in the general areas of reading and word meaning, on the one hand, and mathematics, on the other. (See Table I-2.)

Reading continues to be the outstanding concern of school people so far as school curriculum is concerned, especially in Title I and similar programs, but not far behind is concern for arithmetic achievement. We have gone through many curriculum changes during the period of the last decade. The traumatic experience of a major revision in the mathematics curriculum (from conventional to the so-called "modern" math), especially in the middle grades, is over and currently the trend is back toward a more conventional approach.

Form X of the Intermediate I Battery is intended for grade 4.0 through grade 5.5; in other words, all of the fourth grade and one-half of the fifth grade. All of the items in this test, therefore, should be basically applicable to this grade range - with the possible exception that some very easy items may have been included for the sake of giving "bottom" to the test for the slow learners in grade 4, and some difficult items may have been included in order to give the test "top" for children tested up to the middle of the fifth grade.

The Intermediate I Battery used in this study thus is optimally placed for the designated grade levels (4.0-5.5) and, therefore, a very large proportion of the items should be found within a typical curriculum at grade 4. Just how many of these items are typical of the curriculum in New Hampshire can be told by comparing the items in the test booklet with courses of study available for the state. Variations in the curriculum from school district to school district also are of major importance.

It should be noted here that the representative sample used for this analysis, consisting of some 560+ students, was chosen randomly from all parts of the state and, therefore, any validation that tries to relate the tests to the curriculum in effect in Community A versus Community B is doomed to failure. This may not be a serious matter since the determination of the item content for this battery was done in terms of examinations of textbooks and related materials that were most generally used at the particular grade levels mentioned (4.0-5.5).

In addition to a consideration of the published test, we should realize that this test was preceded by an experimental edition used for item analysis on a large and presumably representative population, and it was only on the basis of the item difficulty and item discrimination values so obtained that the final selection of items was made.

A statement in the Stanford Technical Supplement, which is available for the series, indicates that the intent was to maximize the coverage at grade level by including items with difficulties corresponding to this proportion:

Item Difficulty	Percent of Items For Grade 4.8
80-89	10
70-79	10
60-69	20
50-59	20
40-49	20
30-39	10
20-29	10

Table I-2

Composition of Stanford Achievement Test

Complete Battery - Intermediate I: Form X

Test No.	No. of Items	Test Name	No. of Choices	Time Limits
1	38	**Word Meaning	4	10 min.
2	60	**Paragraph Meaning	4	30 "
3	50	Spelling	4	15
4	61	Word Study Skills	3-4	20
5	122	Language	2-4	41
6	39	**Arithmetic Computation	5*	35 "
7	32	**Arithmetic Concepts	4	20 "
8	33	**Arithmetic Applications	5*	30 "
9	49	Social Studies	4	35
10	56	Science	4	25

* Includes one NG (Not Given) space

** Included in the present study

Composition of Otis-Lennon Mental Ability Test

Elementary II Level: Form J

No.of Items	Test Description	No. of Choices	Time Limits
80	A Measure of Verbal-Educational "G"	5	40 min.

By reference to this table it is possible to see that 60% of the items were supposed to fall generally in the range from 40% passing to 69% passing.

The experimental edition was prepared regardless of the difficulty characteristics of the original items prepared and was, therefore, in a sense a more valid test of the total curriculum for grade 4 than the final test as represented in the Intermediate I Battery: Form X.

Additional information concerning the validity of the test is to be found in the Technical Supplement.

The basic reference on validity appears on page 23 of the Technical Supplement and, quite properly, emphasizes the fact that the validity of the test must be determined essentially in terms of the local curriculum because of the variations in the curriculum from place to place. It also, however, points out that validity, in a general sense, is established by reporting the procedure for determining the content from which items were chosen for inclusion in the test; namely, the analysis of textbooks and related subjects. The specific content of each battery is further defined in Appendix B, which contains item content outlines for most of the subjects.

However, the following quite interesting sentence appears early in this Appendix:

"Furthermore, the Word Meaning, Paragraph Meaning, and Spelling Tests in the upper batteries are of such a nature that Content Outlines are not meaningful for them."

It is not quite clear from the Appendix, and certainly not from this sentence, why the content outlines are not meaningful. Is it that there is so little agreement as to content of reading materials, with respect to vocabulary and type of material, that this cannot be generalized? This seems unlikely.

The usual estimations of grade placement obtained by doing readability indices seem not to have been used for the Paragraph Meaning Test. There is no reference to any sources, such as the Rinsland list or other word lists, to show that the words used were categorized by grades in which they most commonly appear to justify the selection of words used in the Word Meaning Test.

This leaves the local community entirely dependent on its own evaluation of the content for Reading and Spelling - to agree or disagree that it is representative of the material being used as part of the instructional material in reading or in vocabulary development.

The content outlines for the arithmetic tests, on the other hand, are quite specific and very helpful indeed in determining what the content of each test is. When these content outlines are used in connection with the test itself to relate the test to the local curriculum, one cannot go far wrong in determining whether or not these tests measure the objectives of the local curriculum.

One might point out that at the time this test was used in this study in 1969-70 the arithmetic tests probably were more valid than they were at the time they were tried out - because the authors and publisher of the 1964 Stanford found themselves in a dilemma. Modern mathematics was just in the process of being introduced and, anticipating a test lifetime of ten years approximately, one had to anticipate that modern mathematics would become the dominant organizational influence in the math curriculum at the local level. Henceforth, it was essential to provide content that would be satisfying to those who had adopted the modern mathematics while at the same time preparing a test which would be functional in 1964 when the revised test was published.

One word of caution concerning all of the tests in the Stanford Battery, or any other achievement battery, is essential. We have referred to the fact that there was an experimental edition tried out on very substantial numbers of children, carefully selected to be representative of the country as a whole. This experimental edition, naturally, contained items which do not appear in the final edition. These items were eliminated essentially for two reasons:

1. The items proved to be too hard or too easy at the grade levels at which they were tried out;

2. The items proved to be faulty in their construction, i.e. they contained ambiguities or more than one correct answer and, because of these faults, had to be discarded.

The tests included in this study are listed below with the numbers of items in each test and a statement of the item type used, which also is of great importance in considering the validity of the instrument.

Word Meaning - 38 items:

Definition-type introductory statements followed by four choices to correctly satisfy the conditions of the definition.

Paragraph Meaning - 60 items:

Paragraphs, each of which contains two or three completions. The words needed to complete the numbered blank spaces are provided in the form of four choices, only one of which satisfies the demands of the paragraph. This is essentially, then, a four-choice completion-type test.

Arithmetic Computation - 39 items:

Four choices plus NG (Not Given). These items are representative of various phases of arithmetic computation as clearly defined in the Technical Supplement content analyses but better appreciated by an actual study of the booklets themselves. Note in the directions for the test appearing in the booklet that each student is asked to work each example first on scrap paper and then choose the correct response or, if his answer is not there, to mark the NG space. In light of data subsequently available, this is the important consideration.

Arithmetic Concepts - 32 items:

Introductory sentences followed by four choices. This test includes a variety of questions, some of which are hard to subsume under the title "Concepts." For example, the translation of a Roman number to an Arabic number is not really a measure of the extent to which the child understands the Roman numeral and can translate it. It is more a measure of the child's ability to relate one number in the Roman form to its counterpart in the Arabic form. More satisfactory in this respect is an item of the general type indicated in item #10, where the sentence is: "Multiplication is most likely a series of - e. additions, f. subtractions, g. divisions, h. estimations."

Arithmetic Applications - 33 items:

Four choices plus NG. This test is almost strictly analagous to the more conventional arithmetic problem that has been used in the past. Note that the student is supposed to work out his own answer on a separate sheet of scratch paper before marking.

While the above information is helpful in defining the coverage of the test in broad general terms, there is no substitute for a careful examination of the test booklet for Form X and, hopefully, the related material to be found in the Directions for Administering, the Teachers' Guide for Interpretation and Use of Test Results and, most importantly, the Technical Supplement, to which reference is made frequently above.

A General Note on Item Difficulties

It has been clearly stated above that the standard procedure has been used for determining item content for each battery; namely, an analysis of textbooks and related material generally subsumed in the content outlines in categories - with a count of the number of items appearing in the test corresponding to each of these categories.

In actual truth, while this constitutes a very reasonable way of making a test it certainly does not constitute a statement of the materials that one should expect students to master at the stated grade level. In other words, all of the topics covered in all three arithmetic tests of Intermediate I: Form X certainly are not going to be introduced and mastered by all or even a majority of the population of students to be found in grade 4 in our situation.

The Existence or Absence of an Hierarchy

Criterion reference testing, now popular in some quarters, generally must assume an hierarchy in the area of presentation of material, i.e. an order for the introduction of the materials so that knowledges and skills essential for later development are taught and mastered before these new skills are introduced.

Such an hierarchy becomes fairly evident in arithmetic for some or, perhaps, the majority of the topics covered. For example, addition and subtraction must be mastered first in the sense of the pupil having a nearly perfect retention of the 100 addition and subtraction facts and also mastered in the sense that the multiplication tables also are known to the point of near 100% perfect recall as needed. However, as one departs from this simplistic approach to arithmetic computation and gets into other aspects of the content, the hierarchy is not as clear.

Additions of long columns of numbers not only calls on the child to know his number combinations, i.e. the 100 addition facts, but also to hold in mind constantly each new partial sum to which he must add a subsequent number. If the child, for example, is adding ten two-place numbers arranged in columnar form, he must remember eight partial sums before he reaches the final sum of one column. He then must carry everything over a single digit to the adjacent column and proceed with the addition of this column in the same manner in order to get the final sum desired.

It is difficult to place a skill of this nature in an hierarchy, since what is involved is not something that can be totally learned but includes a readiness factor that is more of the general character of mental ability.

In a sense, it may be considered to be comparable to the memory or storage capacity of a computer; if a child doesn't have the capacity for storing and retrieving the information concerning partial sums, the number of such two-place numbers he can add together successfully diminishes rapidly. Such two-place columnar addition is sometimes restricted so seriously that the child's limit may be just adding two two-place numbers, especially if carrying is involved. Other children who have this capacity in excess may add almost a limitless number of two-place numbers without difficulty.

Going beyond arithmetic, however, to reading, spelling, social studies and science, no clearcut hierarchy seems evident at all. Perhaps in beginning reading the knowledge of the sounds of letters and the ability to analyze a word phonetically (and "play back the record," so to speak, to see how a word sounds) and then to compare it with the oral configuration of the word which is in "storage" may constitute the basic characteristics of reading potential.

Instruction in reading, therefore, consists largely of the exposure of the child to large and steadily increasing numbers of words in different combinations. The meanings of these words must be carefully developed together with the hearing configuration and with the difficulties of comprehending them in a continuous passage being emphasized.

Typically, these skills should have been taught to near mastery level by the end of the third grade. Beyond this, the evidence for any hierarchy in reading instruction more or less tends to disappear. An hierarchy, as such, almost completely disappears by grade 6 and, for the most part, very few pupils increase their reading skill (except possibly speed of reading) beyond the level developed by the end of grade 5 - unless they are so apt in reading that it becomes avocational and, thus, generally environmentally developed and not just a matter of exposure within the time allowed for reading within the public schools.

This is neither the time nor the place to develop this concept in detail, but in evaluating the paragraphs included in the Stanford Achievement Test one must look at them from the point of view of whether they are graded in some obvious sense of the word as they move from the simple, short, uncomplicated passages at the beginning of the test to the somewhat more complicated and abstract passages constituting the most difficult parts of the test.

Difficulty values are given in the Technical Supplement, Appendix D, for the various tests included in this study. These difficulty values for the national population were obtained during the February-March period when the standardization program was going on and are reported for the same form of the test used in the study; namely, Form X: Intermediate I Battery.

In this report, another set of item statistics, based on the random sample, are given which agree closely with the statewide data. A similar table is provided for Title I for comparison purposes.

If one will refer forward to these random sample difficulty values in Section II, it will be found to be rather extraordinary how closely the New Hampshire values follow the national pattern, very rarely being more than 10% out of the way in terms of the per cent of children passing the items successfully.

Reliability

Reliability is intended to be a measure of the extent to which the responses on the test are stable from one situation to another. Thus, if one were to give Form X in a given week and follow this by Form Y the next week, one would expect this correlation to be quite high. The sources of the lack of identity of score (more properly defined as rank order) from form to form are not inconsequential. They might be considered basically as follows:

1. The content is not identical. The sample of words used in one word meaning test, or in reading length of the sentences and other characteristics, may vary from the first to the second form, etc., even to a substantial degree. Unless this variation is systematic, i.e. applies with equal or proportional force for all students, the reliability coefficients (inter-form correlations) for these tests would be affected.

Similarly, the item content of the arithmetic tests may not be identical or may not be similarly ordered from form to form. There may be intrinsic difficulties in the separate examples from one form to another form for different individuals taking the test. Any particular example may appear to be an equally good representative of a group of items as another example, but may not in fact be so. All of the items comprising the population of three numbers multiplied by three numbers that could conceivably be contrived will show substantial and stable differences, especially for a particular child. (He may not have learned with equal assurance some addition fact or his multiplication table entry at an earlier stage of his schooling.)

2. In addition, there is one very important source of unreliability; namely, what might be called "quotidien variability"; that is. changes in the child from day to day in the effectiveness of his performance depending on the way he feels, how strongly he is motivated, and other factors tend to cause him to perform differently more or less by chance from one time to another.

This kind of stability in the test, however, can be more easily estimated and the amount of error can be measured in statistical terms and stated as the Standard Error of Measurement, which in this sense would involve only those parts of the variability of measurement attributable to the instability of the child's performance - rather than the characteristics and content of the two forms of the test being compared.

3. Number of items in a test and the distribution of their difficulty values greatly affects reliability, and Stanford subtests tend to be too short.

It should be pointed out, also, that the reliability coefficients as reported in the Stanford manuals are basically maximum values because they make use of 1,000-case random samples from the standardization program, not single communities. Thus, the variability of these populations is nearly as great as the variability of the total group and the reliability coefficients are maximized.

The authors quite appropriately point this out in the Technical Supplement and suggest that the Standard Error of Measurement is perhaps the more stable way of expressing reliability - since the increasing variability is cancelled out when the values required, namely the standard deviation and the correlation coefficient between the two tests, are combined in the appropriate formula.

The above is a rather simplistic approach to the question of reliability since it omits discussion of various methods of obtaining these coefficients, such as the split-half method as compared to the Kuder-Richardson approach and such modifications of the Kuder-Richardson formulas as have grown up in the past few years - one of which is used in the Technical Supplement.

The basic fact remains that one must judge the reliability of the test as being stated in rather absolute terms as reported, and it probably is not too representative of what might happen in a particular community.

What has been said above cannot be construed as a viable criticism of the Stanford Achievement Test if one has read the Technical Supplement and is appreciative of the fact that the reliability coefficients as reported are maximum and that the Standard Errors of Measurement are better statistics to reflect the test's reliability.

Some time has been spent on this discussion of reliability because much use will be made of correlations in this study and correlations among tests are, in turn, greatly affected by reliability of the instruments.

PART II

ANALYSIS OF DATA FOR THE RANDOM SAMPLE FALL 1969 AND SPRING 1970

INTRODUCTION

The earlier Title I Report, entitled "A Description and Evaluation of the Statewide Testing Program in New Hampshire" (1971), to which many references have been and will be made throughout the course of this report, was intended to investigate the extent to which Title I children (with the advantages they had arising from their participation in the special activities and small group instruction characterizing Title I) did any better, relatively, than other children - either similar to themselves in ability and age, or equally atypical of the group or grade represented.

It must be said that in many ways this report raised more questions than it answered, and it very early raised in the mind of this writer the suitability of a general achievement-type measure, such as Stanford Intermediate I: Form X, for the purposes intended; namely, to evaluate growth over a relatively short period of time.

To arrive at this conclusion means rethinking some basic assumptions underlying the CONSTRUCTION of a test such as that used. Incidentally, the repeated use of this test at the beginning and the end of the grade involved (and in this particular study grade 4 only) provided an opportunity for the further analysis of data in new and innovative ways.

Let it be immediately said that the questions raised earlier had nothing to do with the quality of test construction represented by the 1964 Stanford, but simply that an instrument made for one purpose was, perhaps, being unwisely employed for another purpose.

Stanford Intermediate I: Form X was, in its time, a measuring instrument of unchallenged quality for the purpose of arranging children in rank order of their achievement in the various subject matter areas in both a reliable and valid manner and in accordance with the best procedures for test construction available at that time. 1/

Since that time the series has been revised, but no attempt will be made in this report to compare the new test with the old since this would be irrelevant and immaterial.

DISTRIBUTION CHARACTERISTICS

Perhaps it would be well to begin this section by referring to the distributions of raw scores obtained on the Stanford Achievement Test: Intermediate I: Form X in at least two basic areas, Word Meaning and Arithmetic Computation, for the random sample tested fall and spring. The similar distributions for the other three tests included in this study are in the Appendix.

Word Meaning represents a test in which the schools cannot be held totally responsible for the increasing vocabulary from grade to grade - since obviously much of a child's vocabulary, these days especially, comes from the general environment in which he lives.

The impingement of television (and particularly programs like "Sesame Street" and "Electric Company") is very hard to assess in general, but the fact that it does affect the learning of children has been pretty well established. In addition, in the average middle-class or upper-class home there is a very substantial amount of reading material available to children at their own level of development and much additional material is available which is suitable to be read to them. The general environment clearly adds to their mastery of an oral vocabulary, but the extent this is true has never been satisfactorily measured and probably never can be because the variables are too great in number and effect.

As regards the children who come from disadvantaged homes and from environments which do not provide the enrichment mentioned above, one should immediately recognize the atypicality of this situation in the average American scene and allow for it. Nevertheless, since admittance to the public schools is on the basis of age and not vocabulary or intellectual development, these handicapped children do constitute a part of the grade structure at the first, and at any subsequent, grade in the schools of America.

The general policy throughout the country, for years, has been to promote children more or less on the basis of chronological age regardless of achievement, resulting in large numbers of underachieving children at any grade. This unfortunate practice, I think, is giving way to a more rational procedure of attempting to provide a curriculum for each child more or less in terms of his needs or level of development, forgetting grade level; but to say that this has been

---

1/ It would be foolish to disregard the fact that questions were raised about the 1964 Stanford norms. We are talking here not of norms, but of matters of internal validity and reliability.

Answer Sheet Study - II

Raw Score	Cum. %	Sta- nine	Fre- quency	
35	99	9	2	**
34	99	9	1	*
33	99	9	1	*
32	99	9	1	*
31	99	9	3	***
30	99	9	7	*******
29	97	9	6	******
28	96	8	22	**********************
27	93	8	8	********
26	91	8	14	**************
25	89	7	19	*******************
24	86	7	17	*****************
23	83	7	21	*********************
22	79	7	17	*****************
21	76	6	11	***********
20	74	6	22	**********************
19	71	6	26	**************************
18	66	6	36	************************************
17	60	5	31	*******************************
16	55	5	28	****************************
15	50	5	33	*********************************
14	44	5	32	********************************
13	39	4	35	***********************************
12	33	4	25	*************************
11	29	4	21	*********************
10	25	4	24	************************
9	21	3	20	********************
8	18	3	25	*************************
7	13	3	20	********************
6	10	2	22	**********************
5	6	2	17	*****************
4	3	1	5	*****
3	2	1	7	*******
2	1	1	6	******
1	1	1	1	*
			586	

FIGURE II-1
Frequency Distribution, Cumulative Percent Distribution, and Stanines
Plus Histogram Showing Shape of Raw Score Distribution Graphically

RANDOM SAMPLE - WORD MEANING - FALL 1969*

Mean - 15.92 St.Dev. - 7.10

* Each * = one case

Raw Score	Cum. %	Sta-nine	Fre-quency	
37	99	9	4	****
36	99	9	2	**
35	99	9	10	**********
34	97	9	10	**********
33	96	8	14	**************
32	93	8	17	*****************
31	90	7	25	*************************
30	86	7	12	************
29	84	7	23	***********************
28	80	7	28	****************************
27	75	6	24	************************
26	71	6	22	**********************
25	67	6	29	*****************************
24	62	5	30	******************************
23	57	5	31	*********************
22	52	5	29	*****************************
21	47	5	28	****************************
20	42	4	37	*************************************
19	36	4	24	************************
18	32	4	24	************************
17	28	4	22	**********************
16	24	3	18	******************
15	21	3	22	**********************
14	17	3	19	*******************
13	14	3	13	*************
12	12	3	12	************
11	10	2	13	*************
10	7	2	15	***************
9	5	1	10	**********
8	3	1	7	*******
7	2	1	1	*
6	2	1	3	***
5	1	1	3	***
4	1	1	1	*
3	1	1	2	**
2	1	1	0	
1	1	1	1	*
			585	

FIGURE II-2
Frequency Distribution, Cumulative Percent Distribution, and Stanines
Plus Histogram Showing Shape of Raw Score Distribution Graphically

RANDOM SAMPLE - WORD MEANING - SPRING 1970*

Mean - 21.87

St.Dev. - 7.26

* Each * = one case

effected by the present time is to be clearly overly optimistic.

The ungraded primary school was a move in this direction but was never universally adopted. It still represents a very logical place to begin a serious attempt at individualizing instruction, but we have many techniques to develop and substantial change in our ideas about primary grades curriculum before we have achieved even this relatively simple beginning.

In light of this background, let us consider for a moment what we see when we look at the distribution of raw scores on the Word Meaning Test for Stanford Intermediate I: Form X administered in the fall of 1969 to the entire state population at this grade level in New Hampshire. The distribution we will examine (Figure II-1), and others to follow, is not for the entire group but for a random sample, carefully drawn from the whole state, which has been independently shown to be reasonably representative of the state. 1/

In the first place, we find that the distribution is more or less bell-shaped and more or less symmetrical, although it is questionable if it would pass a rigid statistical test of being a normal distribution in the strictly mathematical sense. Existing tests for this purpose are fairly rigorous and, although the number of cases (586 in this particular instance) is fairly large, it is very doubtful if this distribution would be accepted as a random variation from a normal curve if an X^2 test were applied. This really is of relatively small importance.

The curve does show a definite piling up of scores from the middle to the bottom and from the middle to the top, with fewer and fewer children earning very high or very low scores, in a clearly systematic fashion.

The mean of the Word Meaning distribution, as of testing time in the fall (October) of 1969, was 15.9 and the standard deviation was 7.1. The number of items in the test is only 38. The highest score earned in the fall was 35, but there was one case receiving a score of 1!

A test of 38 four-choice items answered purely randomly, without reference to any textual material and without any application of thinking to the marking of the answer spaces, would yield a mean chance score of one-fourth of the total number of items, or 9.5, and a standard deviation roughly equivalent to the number of alternatives, which is 4. In other words, better than 20% of the children taking the test in the fall actually made scores below the mean chance level.

The reported reliability coefficient of this test on an internal consistency basis is better than .90.

The question must arise immediately, however, as to whether this was the appropriate test to use at the fourth grade for the purpose intended; namely, that of measuring the extent of gain, or growth, in the group over a seven-month period by the Title I children as compared to the growth in a random sample for the state - i.e., the population presently under study.

To answer this question, one must look also at the distribution of scores for the spring (Figure II-2). This shows many of the same characteristics, but the slight tendency toward a positive skewness showing up in the fall test now becomes a slightly negative skewness, and the mean score goes from the previously quoted mean of 16, approximately, to 22 - while the standard deviation remains about the same; namely, 7.3 from 7.1. Thus, the raw score gain from fall to spring over seven months in Word Meaning is approximately six points.

Now this is hardly enough average gain to measure with any confidence the gain of individual students. The length of the test, namely 38 items, is obviously too short for the purpose under any circumstances, and the suspicion remains that there is a great deal of guessing involved. Even if this were not so, it would be almost inconceivable that the two curves - i.e., Fall versus Spring - would reflect the same amount of gain for all students, able and retarded.

The fact that the distributions are symmetrical and approach the normal curve strengthens rather than diminishes the hypothesis that guessing is a factor, and a major intent of the present study is to try to assess the effect of such guessing on the scores, both of groups and of individuals, and to make recommendations, finally, as to how an improved type of instrument can be made for the very specialized purpose of measuring gains over a short period of time for all pupils involved.

Perhaps it would be wise at this point to consider the conditions under which a normal curve would arise, assuming all of the marks made by the children taking the test were random and whether random curves

---

1/ See earlier state report entitled "A Description and Evaluation of the Statewide Testing Program in New Hampshire in 1968-69 and 1969-70." (1971)

would reflect any gains over the stated time period.

A random curve would result if one were to hand out answer sheets without test booklets and inform the children that their task was to mark the answers as if they were taking a test. Scoring the test could subsequently be in terms of the established key for the test or on the basis of a random key chosen from a table of random numbers or in some similar fashion.

The writer has done this any number of times as a graduate school exercise, but most recently has asked a colleague, Dr. George Prescott at the University of Maine, to repeat the experiment - using an actual answer sheet for an actual test rather than a standard answer sheet of the IBM type with 150 five-choice responses.

The results were consistent with the writer's earlier studies; namely, that the mean score was equal to approximately that which would be expected by chance and that the standard deviation corresponded closely to the alternatives in the test. 1/

There is NO reason to expect that other than a chance difference would come about if the experiment had been repeated seven months later, however. In other words, the reported gain from fall to spring probably reflects change (growth) from fall to spring as a result of exposure and learning; but is it enough to be actionable or convincing? Is it free enough of guessing to make the results convincing?

Considerations of this sort led the author to think very seriously as to what kind of a test should be used to measure gain or growth over a relatively short period of time, and this constitutes the major purpose of this whole study - especially as it is affected by the factor of guessing and its influence on the nature of the score distribution.

Before continuing, let us next consider the situation with respect to Arithmetic Computation.

The Arithmetic Computation Test of Stanford Intermediate I: Form X contained 39 items as compared to the 38 in Word Meaning. However, the scores ranged from 1 to 29 in the fall, indicating that the test had plenty of top (i.e., was harder) and the initial distribution of raw scores (Figure II-3), if anything, was somewhat more symmetrical. The mean was 11.5 - standard deviation, 4.5.

In the spring (Figure II-4), however, the range of scores was from 3 to 38, which is not surprising. The mean had jumped from 11.5 to 18.3 (about seven points of raw score), while the standard deviation had increased from 4.5 to 7.0, a very significant fact.

These results illustrate the reason why Word Meaning was contrasted in this study with Arithmetic Computation. The results very clearly show the greater effect of in-school learning in the area of Arithmetic Computation as compared to Word Meaning. In Arithmetic Computation, very little incidental learning takes place at home. Programs like "Sesame Street" do not have the impact that they have in vocabulary, and probably very little incidental learning goes on at home in computation because of its specialized nature. Family experience or community living is not that much involved in this area.

In other words, a test intended to measure the outcomes of specific in-school instruction is much more likely to be suitable for the purpose if the content is limited more strictly to the content of the curriculum, as clearly defined in textbook courses of study and particularly the local curriculum, and not much affected by incidental factors.

The distribution of scores in Arithmetic Computation for the spring testing program has the same general symmetrical character as the one for fall, and in both distributions there is an absence of a suggestion of change in skewness from positive to negative - as is evident in the distributions for Word Meaning.

The number of choices in Arithmetic Computation is four numerical options and one option called "NG," or Not Given, which is used sparingly but is definitely used as a keyed response for which credit is given by the authors of the test. It is intended as a kind of escape valve for the pupil who gets a wrong answer by his own computation.

There still is the lingering question, however, as to the extent to which a guessing factor affects the scores, in this instance, in a similar way to that involved in reading.

The standard correction for guessing, which is the number-of-rights less a fraction of the wrongs equivalent to one less than the options offered, has been shown repeatedly to be ineffective and to be totally inoperative if a child answers, or attempts to answer, all of the questions contained in a test.

1/ A summary of this Prescott study is available on request.

Raw Score	Cum. %	Sta- nine	Fre- quency	
29	99	9	1	*
28	99	9	0	
27	99	9	0	
26	99	9	1	*
25	99	9	2	**
24	99	9	0	
23	99	9	6	******
22	98	9	5	*****
21	97	9	4	****
20	97	9	6	******
19	96	8	15	***************
18	93	8	16	****************
17	90	7	24	************************
16	86	7	29	*****************************
15	81	7	34	**********************************
14	76	6	32	********************************
13	70	6	40	**
12	63	5	54	**
11	54	5	48	**
10	46	5	60	**
9	36	4	58	**
8	26	3	44	**
7	18	3	32	********************************
6	13	2	30	******************************
5	8	2	22	**********************
4	4	1	14	**************
3	2	1	7	*******
2	1	1	2	**
1	1	1	1	*
			587	

FIGURE II-3

Frequency Distribution, Cumulative Percent Distribution, and Stanines Plus Histogram Showing Shape of Raw Score Distribution Graphically

RANDOM SAMPLE - ARITHMETIC COMPUTATION - FALL 1969*

Mean - 11.46 St. Dev. -4.47

* Each * = one case

Raw Score	Cum. %	Stanine	Frequency	
38	99	9	1	*
37	99	9	1	*
36	99	9	1	*
35	99	9	1	*
34	99	9	2	**
33	99	9	6	******
32	98	9	8	********
31	97	9	5	*****
30	96	8	19	*******************
29	92	8	11	***********
28	91	7	15	***************
27	88	7	17	*****************
26	85	7	16	****************
25	82	7	12	************
24	80	7	19	*******************
23	77	6	23	***********************
22	73	6	28	****************************
21	68	6	29	*****************************
20	63	6	27	***************************
19	59	5	37	*************************************
18	52	5	36	************************************
17	46	5	23	***********************
16	42	4	30	******************************
15	37	4	28	****************************
14	32	4	28	****************************
13	28	4	34	**********************************
12	22	3	19	*******************
11	18	3	32	********************************
10	13	3	17	*****************
9	10	2	14	**************
8	8	2	13	*************
7	5	1	18	******************
6	2	1	3	***
5	2	1	8	********
4	1	1	2	**
3	1	1	1	*
			584	

FIGURE II-4

Frequency Distribution, Cumulative Percent Distribution, and Stanines
Plus Histogram Showing Shape of Raw Score Distribution Graphically

RANDOM SAMPLE - ARITHMETIC COMPUTATION - SPRING 1970*

Mean - 18.34 St. Dev.-6.97

* Each * = one case

Normative Problems

Since there were no spring norms for Stanford and since such norms would be of doubtful application in any case, the testing of the random sample was inaugurated to provide the very necessary touchstone against which to analyze first a cross section sample of children and then the performance of Title I children. The preservation of the answer sheets for both samples and for both fall and spring made possible the kind of in-depth analysis which we will report in the following pages.

Thoughtful consideration of the problem over a long period of time resulted in a conclusion that this in-depth analysis could be done only by considering each separate item rather than the test score as a whole and, as a result, the data were prepared on IBM cards so as to show the responses made by each child to each item in the tests involved.

THE DIFFICULTY CHARACTERISTICS OF EACH OF THE ITEMS IN THE FIVE TESTS BEING CONSIDERED

In Table II-1, item analysis data are presented for the five Stanford Tests considered in this study for the random sample of 567 students selected for testing in the spring for whom fall test results were also available. The table referred to above presents the data for both Fall and Spring and also presents data separately for Rights, Wrongs, and Omits. Finally, it presents a ratio of the Rights divided by the Attempts (R/A), the significance of which will be discussed in subsequent paragraphs.

Let us consider, first, the percent passing the various items from 1 to N in each test from the point of view of the order of difficulty. Starting with Word Meaning, we see that the items, even in the Fall administration, are generally on the easy side for this sample. No item in the first ten is passed by fewer than 60% of the pupils, and percent passing for most of these beginning items is much higher.

Generally speaking, the authors and publishers of the test put the Word Meaning items in order of difficulty based upon the data from the tryout edition of the test, from which the final forms were made, and it is interesting to see that even after the passage of some years a relatively small group representing a random sample of the fourth grade in New Hampshire shows essentially that this order of difficulty has remained more or less constant - with a surprisingly small number of exceptions.

Perhaps the first ten items of this test, if you consider both Fall and Spring performance, could be considered to be sufficiently mastered at the end of grade 4, so that these words could be considered essentially to be in the working vocabulary of the children - assuming that the percent answering the questions correctly is not too greatly affected by guessing. The criterion used to determine mastery is roughly 75% passing.

In neither Fall nor Spring does a large enough percentage of the group answer the questions correctly from #11 on to permit the assumption that the words in question are in the working vocabulary of the children, and the last half of the test (roughly) contains items of such difficulty that it would be quite unreasonable to suppose that the words were, indeed, part of the working vocabulary of the students involved.

Turning our attention now to Paragraph Meaning and scanning the item difficulty values quickly, especially those for spring, we see that a fair number of items, down to item #13, show a percent passing of .75 or higher; but beyond item #13 there are very few such items and after item #23 the items drop off very rapidly in difficulty or in percent passing.

Paying attention now just to the percentages for spring - that is, at the end of the instructional period - as we move on to Arithmetic Computation, we see the first few items show a fair level of mastery, up through perhaps item #7, and then the items drop off quite rapidly until, after item #14, there are very few items that exceed 50% compared to the total number of items in the test.

For all practical purposes, the last ten items or so in the Arithmetic Computation Test show negligible mastery, on the part probably of the ablest pupils only, so we at this point face up very clearly to the fact that this test is just not suited to the curriculum of New Hampshire, or perhaps it would be better to say it is not suited to the pace with which arithmetic is introduced or the amount of attention paid to it. Certainly if Stanford Computation is to be a guide, the arithmetic situation was serious at the time this test was given.

A word of caution is needed here. This is a test made to measure all levels of ability - not an assessment of a fairly "local" curriculum. A "good" measuring instrument has a mean score at its optimum level of approximately one-half the number of items in the test and the item difficulty values ranging from very low to very high; e.g., .10-.90 possibly. This is why such a test serves so poorly to measure individual pupil gains in a situation like this and hardly serves, even under optimum conditions, as a good measure of group gains.

In Arithmetic Concepts there are very few items overall, from the very beginning of the test, where 75% of the children answered the question correctly in the spring. They can be counted on the fingers of one hand, as a matter of fact.

Looking at this test from the point of view of the criterion reference basic principle of mastery of items in hierarchical form - that is, where a skill at a given level is the basis for a more highly developed skill at another higher level - we see that Arithmetic Concepts completely fails to meet this test.

The performance at the end of the year is typically somewhere in the 50% passing range up to item #26, with generously inter-

Table II-1

Item Difficulty Values
Percent Answering Each Item Right, Wrong, or Omit and Rights/Attempts for Each Item

RANDOM SAMPLE

ITEM No.	Fall Spring	Word Meaning				Par. Meaning				Arith. Comp.				Arith. Concepts				Arith. Appl.			
		R	W	O	R/A	R	W	O	R/A	R	W	O	R/A	R	W	O	R/A	R	W	O	R/A
1	F	.85	.14	.01	.86	.88	.11	.01	[illegible]	.88	.12	.00	.88	.73	.27	.00	.73	.85	.15	.00	.85
	S	.93	.07	.00	.93	.93	.07	.00	.93	.88	.12	.00	.88	.85	.15	.00	.85	.86	.14	.00	.86
2	F	.77	.20	.03	.79	.80	.20	.00	.80	.77	.12	.01	.78	.51	.44	.05	.53	.83	.17	.00	.82
	S	.87	.12	.01	.87	.87	.12	.01	.87	.81	.18	.01	.82	.55	.43	.02	.56	.84	.16	.00	.84
3	F	.76	.22	.02	.77	.83	.16	.01	.84	.77	.22	.01	.78	.69	.31	.00	.69	.73	.26	.01	.74
	S	.88	.12	.00	.88	.93	.06	.01	.94	.85	.15	.00	.84	.75	.25	.00	.75	.76	.23	.01	.77
4	F	.72	.27	.01	.73	.66	.33	.01	.66	.79	.20	.01	.79	.61	.38	.01	.62	.62	.37	.01	.63
	S	.85	.14	.01	.86	.75	.25	.00	.75	.84	.16	.00	.84	.69	.31	.00	.68	.62	.38	.00	.62
5	F	.71	.27	.02	.73	.65	.35	.00	.65	.74	.24	.02	.74	.46	.51	.03	.47	.59	.40	.01	.60
	S	.82	.17	.01	.83	.77	.23	.00	.77	.87	.12	.01	.88	.61	.37	.02	.62	.68	.31	.01	.69
6	F	.60	.38	.02	.61	.68	.31	.01	.68	.61	.37	.02	.62	.33	.63	.04	.35	.36	.62	.02	.37
	S	.72	.27	.01	.73	.81	.19	.00	.81	.85	.15	.00	.84	.38	.59	.03	.39	.51	.48	.01	.52
7	F	.78	.21	.01	.78	.61	.39	.00	.61	.67	.33	.00	.67	.41	.54	.05	.43	.68	.31	.01	.69
	S	.87	.13	.00	.87	.78	.22	.00	.78	.77	.23	.00	.77	.63	.35	.02	.64	.75	.25	.00	.75
8	F	.66	.33	.01	.67	.66	.33	.01	.66	.39	.60	.01	.40	.61	.38	.01	.62	.29	.69	.02	.29
	S	.82	.18	.00	.82	.71	.29	.00	.71	.63	.36	.01	.64	.67	.32	.01	.67	.46	.63	.01	.46
9	F	.65	.32	.03	.66	.68	.30	.02	.69	.46	.53	.01	.46	.31	.67	.02	.32	.61	.37	.02	.62
	S	.79	.21	.00	.79	.80	.19	.01	.80	.65	.34	.01	.66	.50	.49	.01	.51	.66	.32	.02	.67
10	F	.60	.38	.02	.61	.75	.24	.01	.76	.45	.53	.02	.46	.43	.54	.03	.45	.29	.65	.06	.32
	S	.75	.24	.01	.75	.81	.19	.00	.81	.65	.34	.01	.65	.45	.54	.01	.46	.48	.49	.03	.50
11	F	.49	.48	.03	.51	.77	.23	.00	.77	.23	.59	.18	.28	.53	.47	.00	.53	.31	.62	.07	.33
	S	.69	.29	.02	.70	.83	.16	.01	.84	.58	.37	.05	.61	.53	.47	.00	.53	.45	.52	.03	.46
12	F	.43	.53	.04	.45	.41	.58	.01	.41	.29	.65	.06	.31	.48	.48	.04	.51	.61	.36	.03	.63
	S	.59	.40	.01	.60	.57	.42	.01	.58	.65	.32	.03	.67	.59	.40	.01	.59	.67	.33	.00	.67
13	F	.73	.23	.04	.76	.66	.33	.01	.66	.41	.51	.08	.44	.17	.80	.03	.18	.41	.55	.04	.44
	S	.83	.15	.02	.84	.75	.24	.01	.76	.71	.27	.02	.73	.33	.65	.02	.33	.53	.46	.01	.53
14	F	.57	.37	.06	.60	.55	.44	.01	.56	.35	.58	.07	.37	.57	.42	.01	.58	.49	.49	.02	.50
	S	.65	.34	.01	.66	.70	.29	.01	.70	.53	.44	.03	.55	.67	.33	.00	.67	.57	.42	.01	.57
15	F	.43	.48	.09	.47	.38	.61	.01	.38	.35	.58	.07	.38	.53	.45	.02	.53	.41	.56	.03	.42
	S	.65	.33	.02	.67	.46	.53	.01	.46	.48	.49	.03	.49	.63	.36	.01	.63	.47	.52	.01	.42

Table II-1, Page 2 - Item Difficulties, % Right, Wrong, or Omit and Rights/Attempts - Random Sample

ITEM No.	Fall Spring	Word Meaning				Par. Meaning				Arith. Comp.				Arith. Concepts				Arith. Appl.			
		R	W	O	R/A	R	W	O	R/A	R	W	O	R/A	R	W	O	R/A	R	W	O	R/A
16	F	.63	.31	.06	.67	.71	.28	.01	.71	.27	.66	.07	.29	.22	.69	.09	.24	.59	.38	.03	.61
	S	.74	.24	.02	.76	.79	.21	.00	.79	.51	.46	.03	.52	.34	.61	.05	.36	.70	.29	.01	.70
17	F	.59	.32	.09	.65	.31	.68	.01	.32	.39	.52	.09	.42	.52	.46	.02	.53	.51	.44	.05	.53
	S	.69	.27	.04	.72	.42	.57	.01	.42	.54	.42	.04	.56	.64	.35	.01	.64	.59	.40	.01	.59
18	F	.41	.47	.12	.47	.63	.35	.02	.64	.32	.43	.25	.43	.37	.55	.08	.40	.47	.48	.05	.49
	S	.57	.40	.03	.59	.76	.23	.01	.76	.65	.27	.08	.71	.61	.37	.02	.62	.62	.37	.01	.62
19	F	.39	.44	.17	.46	.57	.40	.03	.58	.37	.50	.13	.43	.55	.41	.04	.56	.40	.55	.05	.42
	S	.61	.35	.05	.64	.71	.28	.01	.72	.59	.37	.04	.61	.61	.38	.01	.62	.55	.43	.02	.56
20	F	.45	.37	.18	.55	.43	.54	.03	.44	.25	.57	.18	.31	.17	.76	.07	.19	.25	.65	.10	.27
	S	.59	.36	105	.62	.58	.41	.01	.58	.43	.50	.07	.47	.30	.67	.03	.31	.52	.46	.02	.53
21	F	.43	.35	.22	.55	.66	.31	.03	.68	.08	.56	.36	.12	.43	.52	.05	.45	.46	.46	.08	.51
	S	.55	.40	.05	.57	.77	.22	.01	.77	.48	.41	.11	.53	.53	.46	.01	.53	.59	.39	.02	.60
22	F	.33	.40	.27	.45	.66	.31	.03	.70	.17	.41	.42	.29	.60	.34	.06	.64	.33	.54	.13	.37
	S	.47	.45	.08	.51	.77	.22	.01	.78	.33	.39	.28	.45	.69	.30	.01	.69	.43	.52	.05	.45
23	F	.32	.38	.30	.45	.53	.44	.03	.54	.15	.51	.34	.24	.45	.46	.09	.49	.23	.62	.15	.26
	S	.49	.43	.08	.53	.61	.38	.01	.61	.46	.40	.14	.53	.57	.39	.04	.60	.39	.57	.04	.40
24	F	.34	.34	.32	.50	.29	.65	.06	.31	.17	.48	.35	.27	.33	.55	.12	.37	.11	.71	.18	.14
	S	.60	.31	.09	.66	.39	.59	.02	.40	.38	.47	.15	.44	.41	.55	.04	.43	.26	.69	.05	.27
25	F	.35	.30	.35	.54	.26	.67	.07	.28	.07	.47	.46	.12	.41	.45	.14	.48	.24	.57	.19	.30
	S	.62	.29	.09	.68	.28	.69	.03	.29	.23	.44	.33	.35	.49	.47	.04	.50	.23	.71	.06	.25
26	F	.37	.25	.38	.59	.19	.72	.09	.22	.13	.40	.47	.25	.43	.42	.15	.51	.21	.59	.20	.26
	S	.61	.29	.10	.68	.28	.68	.04	.29	.34	.46	.20	.43	.63	.33	.04	.66	.33	.60	.07	.35
27	F	.21	.34	.45	.38	.43	.49	.08	.47	.09	.39	.52	.18	.13	.68	.19	.16	.17	.61	.22	.21
	S	.40	.46	.14	.47	.41	.57	.02	.42	.18	.46	.36	.28	.23	.70	.07	.25	.27	.66	.07	.28
28	F	.19	.32	.49	.38	.38	.53	.09	.42	.13	.38	.49	.26	.43	.38	.19	.54	.04	.76	.22	.05
	S	.39	.43	.18	.48	.52	.46	.02	.53	.43	.32	.25	.58	.59	.35	.06	.62	.17	.76	.07	.18
29	F	.09	.41	.50	.18	.42	.48	.10	.43	.09	.39	.52	.18	.13	.61	.26	.17	.09	.65	.26	.12
	S	.30	.52	.18	.37	.54	.43	.03	.56	.29	.43	.28	.40	.22	.68	.10	.25	.21	.68	.11	.24
30	F	.21	.22	.57	.50	.31	.57	.12	.36	.07	.39	.54	.16	.05	.73	.22	.07	.24	.47	.29	.34
	S	.37	.37	.26	.49	.43	.55	.02	.44	.24	.44	.32	.35	.17	.76	.07	.18	.33	.57	.10	.37

Table II-1, Page 3 - Item Difficulties, % Right, Wrong, or Omit and Rights/Attempts - Random Sample

ITEM No.	Fall Spring	Word Meaning				Par. Meaning				Arith. Comp.				Arith. Concepts				Arith. Appl.			
		R	W	O	R/A	R	W	O	R/A	R	W	O	R/A	R	W	O	R/A	R	W	O	R/A
31	F	.17	.24	.59	.40	.41	.45	.14	.47	.09	.35	.56	.20	.13	.60	.27	.18	.20	.48	.32	.29
	S	.38	.35	.27	.52	.51	.47	.02	.52	.27	.38	.35	.41	.21	.69	.10	.23	.33	.56	.11	.37
32	F	.13	.23	.64	.38	.47	.38	.15	.55	.05	.43	.52	.11	.19	.54	.27	.26	.09	.56	.35	.14
	S	.29	.39	.32	.43	.62	.34	.04	.64	.25	.44	.31	.37	.23	.66	.11	.26	.15	.71	.14	.18
33	F	.15	.20	.65	.41	.35	.48	.17	.42	.07	.34	.59	.17					.09	.55	.36	.15
	S	.38	.31	.31	.55	.45	.51	.04	.46	.22	.39	.39	.37					.16	.68	.16	.19
34	F	.13	.19	.68	.38	.45	.35	.20	.56	.05	.34	.61	.13								
	S	.27	.35	.38	.43	.64	.31	.05	.67	.07	.51	.42	.13								
35	F	.08	.21	.71	.27	.37	.42	.21	.46	.06	.36	.58	.14								
	S	.25	.37	.38	.40	.53	.41	.06	.56	.16	.43	.41	.27								
36	F	.05	.21	.74	.20	.45	.31	.24	.59	.03	.35	.62	.09								
	S	.16	.44	.40	.27	.61	.32	.07	.65	.10	.44	.46	.19								
37	F	.06	.18	.76	.26	.42	.32	.26	.57	.05	.34	.61	.13								
	S	.22	.34	.44	.39	.65	.27	.08	.70	.13	.39	.48	.25								
38	F	.05	.20	.7[illegible]	.18	.41	.32	.27	.56	.05	.28	.67	.17								
	S	.13	.42	.45	.25	.62	.28	.10	.68	.11	.38	.51	.22								
39	F					.31	.39	.30	.44	.09	.28	.63	.23								
	S					.52	.37	.11	.58	.13	.37	.50	.26								
40	F					.39	.30	.31	.56												
	S					.57	.32	.11	.64												
41	F					.21	.45	.34	.31												
	S					.39	.50	.11	.43												
42	F					.37	.24	.39	.61												
	S					.59	.26	.15	.70												
43	F					.35	.25	.40	.58												
	S					.59	.25	.16	.70												
44	F					.25	.32	.43	.43												
	S					.43	.40	.17	.52												
45	F					.27	.27	.46	.49												
	S					.47	.33	.20	.59												

Table II-1, Page 4 - Item Difficulties, Random Sample

ITEM No.	Fall Spring	Par. Meaning R	W	O	R/A
46	F	.17	.36	.47	.31
	S	.35	.46	.19	.43
47	F	.10	.40	.50	.20
	S	.23	.55	.22	.29
48	F	.27	.23	.50	.54
	S	.47	.29	.24	.61
49	F	.13	.30	.57	.29
	S	.23	.49	.28	.33
50	F	.05	.37	.58	.11
	S	.10	.60	.30	.14
51	F	.18	.21	.61	.05
	S	.35	.34	.31	.50
52	F	.12	.25	.63	.33
	S	.33	.35	.32	.48
53	F	.19	.17	.64	.52
	S	.42	.25	.33	.62
54	F	.11	.24	.65	.30
	S	.26	.39	.35	.40
55	F	.15	.20	.65	.44
	S	.31	.33	.36	.48
56	F	.16	.18	.66	.48
	S	.35	.26	.39	.57
57	F	.06	.26	.68	.19
	S	.15	.44	.41	.25
58	F	.09	.22	.69	.30
	S	.23	.36	.41	.38
59	F	.07	.22	.71	.24
	S	.11	.44	.45	.21
60	F	.09	.20	.71	.31
	S	.17	.38	.45	.32

spersed higher values for a few items before this, but after item #26 almost nothing is shown that indicates even understanding, let alone mastery. The figures reported could be actually the result of chance.

In Arithmetic Applications there are three items in the beginning of the test that show a high level of mastery, but the subsequent difficulty values begin then to fall off precipitously almost immediately. Item #7 reaches 75%, but it stands out as being very much the exception.

Continuing on through the test, the general trend is for items to be answered in the 50% to 60% range down to about item #22, after which there is another precipitous fall with as few as 17% answering item #28 correctly. Here, certainly, many of the items are measuring things that have not been presented to the group formally or taught in any real sense of the word. It is the writer's best guess that the performance here, while it looks fairly good, is largely the result of the ability of the ablest students to handle the arithmetic situation "on their own."

In all of this discussion, especially of the Arithmetic Tests, a person reading this study should have before him the test booklet itself - so that he can see exactly the kinds of items that children were able or unable to answer in the spring of 1970 and ask if this is a reasonable situation. In other words, was the Stanford Test so far out of line with the New Hampshire curriculum that it never should have been used at this grade level?

Table II-2

Correlations in Raw Scores
Between Otis-Lennon and Selected Stanford Tests

RANDOM SAMPLE - Grade 4 - Fall 1969

	Raw Score Correlations of Otis-Lennon with Stanford		
Selected Stanford Form X Tests	NH Data Grade 4	Data from Otis Manual Grade 3	Grade 5
Word Meaning	.72	.62	.77
Paragraph Meaning	.73	.60	.78
Arithmetic Computation	.42	.50	.60
Arithmetic Concepts	.65	.67	.73
Arithmetic Applications	.60		.75

Item Performance vs Normative Interpretation

Remember now, we're talking about individual items. Factors such as overall - i.e., average - difficulty, norms, rank order, and so forth, are of no significance. The number of cases, amounting to 567 students, has been shown to be generally comparable to the whole state. It is large enough so that the errors of measurement in these percentages are small.

We must therefore, in retrospect, determine whether we can at all be satisfied with the arithmetic performance of New Hampshire students if these data truly represent what they are able to do, especially considering the fact that these are mostly five-choice multiple choice questions and even the percentages as reported are inflated due to the number of correct responses which are correct sheerly by random marking.

We have not said anything, as yet, about the number of omitted items. Actually, in an ideal situation a child should mark only the items he knows and omit the rest. Let us say that it is considered permissible to make an intelligent guess in a four- or five-choice item (Word Meaning being four). This would account for few additional "Rights" due to "guesstimation"; that is to say, partial knowledge is used positively. Those children who have to guess on the meaning of the word certainly would not be qualified as being masters of the word with regard to its use in general conversation or in writing.

Yet it must be emphasized repeatedly that the content of this test was taken from sources which indicated they were generally recognized to be suitable for use in the fourth grade. Naturally, the words in the total test have to cover a wide range of difficulty because the teacher has to cope with a wide range of ability, whether this is desirable or not, and this test was intended as a measuring instrument.

Such a statement can be strengthened by relating the Word Meaning data from the Stanford Achievement Test to information from the so-called intelligence test or mental ability test. In this particular instance, the Otis-Lennon Mental Ability Test was used and the results of its use are reported in the aforementioned Title I Report. 1/ To amplify this we are including here Table II-2 giving the correlations of Otis-Lennon with the five Stanford Tests we are investigating for our own group plus comparable sets of similar data for other groups. 2/

Many people argue that the Otis-Lennon Test is, after all, essentially another vocabulary test - not too different from the vocabulary (Word Meaning) test in the Stanford Achievement Test. The relevance of this comment is pertinent to our problem. However, the Otis-Lennon Test measures far more than just vocabulary - including (as it does) arithmetical problems, spatial reasoning problems, analogies, and a whole variety of mental skills and knowledges that are not specifically curriculum oriented.

It makes little difference whether the skill demonstrated on the Otis-Lennon or other similar mental ability tests arises from native intelligence, i.e. inherited mental ability, or from a good or poor environment - whatever that might be. Whatever it is, quality of environment is not to be measured in terms of dollars and cents of salary earned by the parents of the child or children in question. This has been repeatedly shown to be a fallacy in individual cases, even though there is a positive correlation as shown by group-type analysis. (See data from the Metropolitan Manual for Interpreting, Revised 1972, concerning the relationship of mental ability to socio-economic status - e.g., salary of parent? education of parent? - in the standardization groups for this battery.)

It may appear to strengthen the argument of the environmentalists to note that it can be easily shown that not every word in the Stanford Word Meaning Test occurs in the curriculum for every school (or most schools) in the United States at grade 4 or even the adjacent grades of 3 and 5. On the other hand, analysis of the words that are included in the Stanford Achievement Test: Form X for the Intermediate I Battery shows that they represent a good cross section of words occurring in the kinds of children's literature to which the average child in an average family is exposed at this level of development.

The "curriculum validity" problem really arises from an unrealistic desire on the part of school people and, more particularly, parents and the public in general to have children master everything presented to them within the walls of the school at the grade levels specified. This is totally unreasonable in the case of Word Meaning, especially in view of the conditions as they presently exist, and there is ample statis-

---

1/ Page 19, Table III-B-2

2/ Grades 3 and 5 correlations are from the 1969 Otis-Lennon Mental Ability Test Technical Handbook.

tical and common sense evidence to establish this point. Arithmetic may be an entirely different matter, since environmental learning is much less effective here.

What then can we say about the Stanford Achievement Test: Word Meaning: Intermediate I Battery: Form X as an instrument suitable for the purpose for which it was used; namely, to measure achievement in vocabulary at the beginning and end of grade 4?

As a measuring instrument, it has served the purpose well. In other words, it has selected those individuals who have a high vocabulary and has similarly identified those who have a paucity of skill in that area. This is very valuable information for the teacher and is quite irrelevant to the specific words which may be taught at the local level.

As a matter of fact, there are few situations where vocabulary, as such, is taught independently of the total language program, which includes reading, speaking, spelling, and the use of the English language in writing.

On the negative side, the Stanford Word Meaning Test is quite obviously too short, and therefore too limiting in proportion of words which will be found in a local curriculum, to measure specific outcomes of even the most carefully planned "new" programs of instruction. Children will not have been exposed in a specific learning situation to a great number of these words, but will have learned them quite incidentally both in their schoolwork and in the home and community in general. A radical solution to the problem may be necessary, and in due time in this report we will attempt to attack that problem.

In the meantime, it is essential that we turn our attention to the comparisons between the percent of items answered incorrectly and the percent of items omitted. What we find here is that the percent of items answered incorrectly is not too different from the percent answered correctly, except for the very easy or very difficult items, and the percent of answers omitted is substantially small. In other words, children are marking answers in far greater proportion than they would if you, i.e. the teacher or the school, expected them to mark only those words where they felt they had a reasonable chance of really knowing the word. In relatively few cases are they actually omitting items in large number; therefore, the case for random guessing is greatly strengthened and the validity of the test for measuring anything is weakened.

Let us follow up a little more closely the suggestion just made. The writer may report in this connection a fairly large number of instances where he has queried children individually concerning their test-taking behavior. Almost uniformly, the response was that they view a multiple choice question (or any of its variants) as simply a situation where they answer the questions immediately, i.e. perceptively, if they know what the answer is.

If they do not know, they canvass the possible right answers as given and choose the one that seems to be the most likely and mark it. If they can find no clues as to what the correct answer is among the words provided as alternatives, they simply mark an answer by chance in the hope of getting an unearned credit, at least until they recognize that they are simply beyond their depth. Even then, a remarkable number just continue to mark all answers in the test.

The question for further study is, "Is this what children actually do?" The data to be reported later will reveal the extent to which this appears to be the case.

ANALYSIS OF PUPIL RESPONSES BY CATEGORY

One unique bit of information that is available is the result of the fact that we do have fall-spring item analysis data showing the response of each pupil to the identical items on two occasions. The responses are separated by a period of approximately seven months. Thus we are able to determine the consistency (or lack of consistency) in the pupil responses over a period of learning covering the better part of the school year.

One of the first methods of attack was to create categories of response which would describe how a pupil had answered an item in the fall versus the spring when these two periods were considered jointly.

An example of this type of categorical analysis is the "RR" (Right in the fall, Right in the spring) category. An item falling in this category would be totally useless for measurement of learning resulting from a particular program of instruction since it would simply demonstrate that the learning that had taken place prior to the testing time in the fall was maintained through the period of seven months.

The individuals who were involved responded to the item correctly even after this passage of time, barring the quite remote chance of fortunate guessing fall and spring. Result: teaching effort is wasted.

The existence of such items in effect reduces the length of the test as a measuring instrument, the representativeness of its coverage, and its reliability and validity - whether this test be Word Meaning, or Paragraph Meaning, or Arithmetic.

A logical analysis of the possible categories reveals that the ten decided upon would almost exclusively cover every possible response a pupil might make to an item within the established response framework; i.e., multiple choice with answer sheet.

All pupil-item responses (number of pupils times number of items) are broken down by category and presented in two tables. The two tables overlap in that the numbers of pupil-item responses involved in each category are repeated, but in one table are interpreted in terms of a mean per category and in the other table, in terms of a percentage per category.

Interpretation in Terms of Mean Per Category

Let us consider first Table II-3, in which a value therein labeled "Mean Responses" is presented below the number of pupil-item responses in each category. These mean values were found, for example, by dividing the number of pupil-item responses under the category "RR" by the total number of pupils, which in the random sample was 567 cases including both boys and girls.

(Actually boys and girls were studied separately, but no significant sex differences were found and, therefore, for this report the data are combined.)

When this process is carried out, the quotient is the average number of test items falling in that category for the group tested.

The results for all of the categories are interesting in that each reveals one thing or another. For example, the "WWS" (Wrong in the fall, Wrong in the spring, Same choice) category would suggest that a pupil or a number of pupils might have had some positive misinformation which was preserved over the period of time during which they were under instruction; while the "WWD" (Wrong, Wrong, Different) category almost surely identifies those who did not answer the question on a basis of specific knowledge at all, but merely marked a response by chance.

Similarly, the "OO" (Omit, Omit) category represents the children who refused to commit themselves, either fall or spring, in a situation where they felt no competency. They are temperamentally "no guessers."

At this moment, however, we're concerned with two response categories which can be readily combined; namely, the "WR" (Wrong, Right) and the "OR" (Omit, Right) responses. Only in the case of these two categories can we concede that learning most likely has taken place as evidenced by the test results since only in these categories do we find that an initial response, which indicates that "learning" or "mastery" has NOT previously taken place, has changed to a response which indicates that now the pupil may, indeed, have learned the answer to the questions involved; i.e., to answer a question which he was previously unable to answer.

Continuing now with Word Meaning, for the sake of further illustration, when the "WR" and "OR" categories were added together for the random sample, the total number of pupil-item responses was 5,053. When 5,053 is divided by 567, the result is the average number of items answered in a manner to suggest an increment in mastery of the material in question - in this case vocabulary - during the seven-month period. This gives a mean number of items on which learning has probably taken place of 8.9.

Table II-3

Number and Mean of Pupil-Item Responses by Consistency Category

RANDOM SAMPLE of 567 Boys and Girls

Test	No. of Items	Possible Fall-Spring Responses	MARKED RESPONSES*					OMITS**				
			RR	RW	WR	WWS	WWD	OO	OR	OW	RO	WO
Word Meaning	38	21,546	7298	1483	3205	1413	1600	1927	1848	2132	211	429
Mean Responses			12.9	2.6	5.6	2.5	2.8	3.4	3.3	3.8	.4	.7
Paragraph Meaning	60	34,020	10100	3281	5482	2619	3476	2950	2516	2709	309	578
Mean Responses			17.8	5.8	9.7	4.6	6.1	5.2	4.4	4.8	.6	1.0
Arithmetic Computation	39	22,113	4716	1552	4155	1522	2667	2608	1475	2220	229	969
Mean Responses			8.3	2.8	7.3	2.7	4.7	4.6	2.6	3.9	.4	1.7
Arithmetic Concepts	32	18,144	5159	2078	3533	3183	2314	208	526	788	84	271
Mean Responses			9.1	3.7	6.2	5.6	4.1	.4	.9	1.4	.1	.5
Arithmetic Applications	33	18,711	4973	2177	3582	2255	3289	365	576	1120	93	281
Mean Responses			8.8	3.8	6.3	4.0	5.8	.6	1.0	2.0	.2	.5

*RR-Right Fall and Spring
RW-Right Fall-Wrong Spring
WR-Wrong Fall-Right Spring
WWS-Wrong Fall and Spring-Same Response
WWD-Wrong Fall and Spring-Different Response

**OO-Omit Fall and Spring
OK-Omit Fall-Right Spring
OW-Omit Fall-Wrong Spring
RO-Right Fall-Omit Spring
WO-Wrong Fall-Omit Spring

Note particularly that this does not identify the particular words which have been learned, and that these words may not indeed be the same from pupil to pupil; it simply emphasizes the fact that out of 38 items, a total population of 567 came up with an average of 9 items which appear to have been learned during the seven months.

A moment's thought makes it clear that this line of reasoning cannot be followed in a single testing. Any fall Wrong or Omit can be transformed to a Right response in the spring because of real learning. Only the opportunity provided by the fall-spring analysis reveals the small average number of items learned. Similarly, some of the "RR" responses do not really reveal positive learning - because both "Rs" may have come about by guessing, a real but remote possibility.

What is lacking, therefore, is prior assurance of a serious effort to test what the teacher teaches during the seven months in question - without encouraging "teaching for the test." This "community" curriculum is only approximately "knowable" beforehand for any standardized test, and there is no infallible way of freeing the teaching situation of the totally undesirable effect of the "coaching" dilemma.

An ideal test would be one with a large number of responses "Wrong" in the fall, all of which were previously certified locally as valid teaching objectives during the coming year. Items not taught, but learned anyway, give false credit to the school; items taught, but not learned, raise questions about the effectiveness of instruction.

Subsets of locally valid items may be selected from standardized tests by an appropriate local (logical) analysis of the test items based upon the established goals for the year - a long-recommended practice. However, a desirable practice becomes a required practice if the intent of testing is specifically the evaluation of local teaching efforts.

This conclusion is obvious enough but is differently stated when one says, as above, that only the Wrong, Right or Omit, Right items can provide evidence of growth. The way to demonstrate more growth is to make a special variety of test by which only the items taught are considered in determining changes attributable to the child's instruction. Obviously additional determining factors are the level of motivation in taking the test coupled with freedom from guessing. Not guessing by choice because one wishes to be honest - i.e., to reveal his areas of ignorance as well as knowledge - is an outcome of a good teacher-pupil relationship.

Consider now, by way of reenforcement of the above, the fact that all categories except "WR" and "OR" are in a sense "disabled" - in that they cannot reveal that any learning has taken place.

If a child answers a question "RR," this simply means that he knew something at the beginning and continued to know the answer at the end of the period of instruction. A "WWD" response is highly suggestive of guessing; etc. If only 9, or less than one-quarter, of the questions show average positive change over seven months, the test obviously cannot possibly be analytical for an individual child.

Unfortunately, all of the circumstances involved in the collection of these data suggests that the instrument was not an appropriate one to prove the effectiveness of instruction in the field of vocabulary development with this population. Any survey instrument, excellent though it is for the purpose intended, cannot be of sufficient effective length to establish curriculum validity for the individual school administrative units involved.

We turn now to Arithmetic Computation, in which most learning actually takes place in the school and not in the general environment. The average is 9.9, or about 10 items or 10 learnings resulting from the seven-month period of instruction. (The two tests are specifically chosen to provide a contrast because one is so obviously influenced by the general environment and the other one is not so obviously influenced by this environment.)

Note that in both of the instances quoted above we are talking about averages. These are arithmetic means and, therefore, no statement can be made concerning the percent of children learning more or less than the mean - unless we can further assume that the distributions are symmetrical, in which the mean and the median would be the same.

The measurement of short-term gains is difficult indeed and is doomed to be inconclusive or ambiguous unless one can establish that the knowledge involved was not known at the beginning of instruction and was mastered by an established percent of individuals at the end of the period of instruction. Considering variations in the Title I projects submitted and looking also at the wide range of achievement and ability of a group of students in any typical class, the situation is even more complicated!

It is also perfectly evident that we must have some assurance that the pupil group involved in the experiment is able; that is, ready to learn what the LOCALLY VALID test measures.

We also must be assured that instructional time allowed will be sufficient. We can assume about 180 days of in-school time per year, or about 140 days in seven months between first, or fall, and second (spring) testing time, but the minutes allowed per day are variable, both from subject to subject and unit to unit.

We can guess that the total amount of time involved in actual vocabulary development, including or involving the particular words in the SAT Word Meaning Test, probably would be small; but there are other factors involved, such as incidental outside word or vocabulary learning, which make this a bad subject for evaluative purposes.

If the in-school instruction had as its main purpose the development of widely applicable methods of word attack, the particular subpopulation of words in the test would not be as important. A pupil could apply these skills to answering any Word Meaning items - a desirable goal but one we cannot assume was characteristic of our population.

Let's turn our attention now to what is true of the Arithmetic Computation Test, where we can tie down much more definitely what learning tasks are facing the pupils of grade 4 during the seven-month period under investigation if they are to cope adequately with the Stanford Arithmetic Tests.

If we assume the same 140 days of time and an allotment of one-half hour per day to instruction in arithmetic, with a major emphasis at this grade level on computation, we come up with a total of about 70 hours of instruction over the seven months. Is this enough?

Perhaps our estimate per day is too low. What if we assume 60 minutes? Would that be enough? It would be a viable project to see what would happen, comparatively, if 50% to 100% more time were allowed, or if a small amount of time per day were devoted to maintenance of skills in oral arithmetic.

If we further assume that this instruction was carried on in the average self-contained classroom with its typical wide spread of talent, it is probably unlikely that more than half of the members of such heterogeneous classes ever could really master any except the simplest of the knowledges to which they are theoretically exposed but which they did not partly know when the test was first administered. What then?

In point of fact it is horrendous, from a scientific point of view, to draw conclusions in any subject field without knowing and stating these facts. God forgive us for what we do in the name of educational evaluation!

In defense of the instrument involved (and of testing in general), it must be remembered that the content of the test was taken directly from the typical content in arithmetic computation texts for grades 4 and 5. The assignment of text content to grade is not a matter of 100% agreement, even in arithmetic!

In other words, there is no hard and fast hierarchy that says that "A" must be learned before "B" and "B" learned before "C" even in arithmetic - or even, more particularly, in arithmetic computation. Hence an item which might be a fourth grade item in one system or one curriculum might be assigned to the third or fifth grade in another curriculum, etc.

This simply means that the content of the test must be defined in terms of the curriculum arrived at by the agency which is responsible for making such curriculum decisions - whether this is the local community, the county, or the state.

In New Hampshire (where this experimentation was carried on), theoretically at least, the decisions usually are made at the school district level or lower, without any really notable interference at the state level - although the State Department of Education exercises some influence in determining desirable objectives, especially in fields as specific as arithmetic computation. There is no mandated textbook in any subject and no set course of study to which all must adhere.

Interpretation of the Ten Categories in Percents

In Table II-4 the same pupil-item data are presented, but the method of interpretation used is different. It is intended to reflect the proportion of all possible pupil-item responses, or interactions, that suggest that learning has taken place as compared to the total number of such pupil-item responses included in the test, category by category.

The same argument given above holds here. The only categories unequivocally revealing positive changes in the direction of

Table II-4

Number and Percent of Pupil-Item Responses by Consistency Category

RANDOM SAMPLE of 567 Boys and Girls

Test	No. of Items	Possible Fall-Spring Responses	MARKED RESPONSES*					O M I T S**				
			RR	RW	WR	WWS	WWD	OO	OR	OW	RO	WO
Word Meaning	38	21,546	7298	1483	3205	1413	1600	1927	1848	2132	211	429
Percent of Responses			33.9	6.9	14.9	6.5	7.4	8.9	8.6	9.9	1.0	2.0
Paragraph Meaning	60	34,020	10100	3281	5482	2619	3476	2950	2516	2709	309	578
Percent of Responses			29.7	9.6	16.1	7.7	10.2	8.7	7.4	8.0	.9	1.7
Arithmetic Computation	39	22,113	4716	1552	4155	1522	2667	2608	1475	2220	229	969
Percent of Responses			21.3	7.0	18.8	6.9	12.1	11.8	6.7	10.0	1.0	4.4
Arithmetic Concepts	32	18,144	5159	2078	3533	3183	2314	208	526	788	84	271
Percent of Responses			28.4	11.5	19.5	17.5	12.8	1.1	2.9	4.3	.5	1.5
Arithmetic Applications	33	18,711	4973	2177	3582	2255	3289	365	576	1120	93	281
Percent of Responses			26.6	11.6	19.1	12.1	17.6	1.9	3.1	6.0	.5	1.5

*RR-Right Fall and Spring
RW-Right Fall-Wrong Spring
WR-Wrong Fall-Right Spring
WWS-Wrong Fall and Spring-Same Response
WWD-Wrong Fall and Spring-Different Response

**OO-Omit Fall and Spring
OR-Omit Fall-Right Spring
OW-Omit Fall-Wrong Spring
RO-Right Fall-Omit Spring
WO-Wrong Fall-Omit Spring

learning are the "WR" and "OR" categories. When the number of cases, i.e. pupil-item responses, in these two categories are combined and this number is divided by the total possible number of such pupil-item responses (which varies, of course, from test to test) the results show a remarkable consistence.

The percent of such pupil-item responses which appear to fall in the probable learning category is 22% to 25%. In other words, 25% or less of the possible pupil-item responses indicate that learning did, in fact, take place.

In view of the four- or five-choice multiple choice nature of the present material, we need to be acutely aware of the "RW" and "RO" responses - which suggest the fall Right responses were the result of guessing in the fall.

If a teacher is operating on the basis of fall data, she may be misled by the fall response of those falling in the "RW" and "RO" categories. Some fall responses are probably guesses if the "RW"-"RO" data can be credited. In other words, "money in the bank" by the fall performance was not there! Obviously, item analysis data are also invidiously affected.

(The "RW + RO" and "WR + OR" data are summarized in Table II-5.)

Thus we must conclude that the analytical response approach has the virtue of alerting us to an often sensed but rarely documented fact that item analysis data can be misleading if based on a single measure.

Perhaps a comment is in order concerning the guessing (or forgetting) that does take place among those who mark an item "RW" or "RO." Such "Right-in-the-fall versus Wrong-in-the-spring" responses are particularly vexing because the fall item analysis of Rights is so misleading. Nineteen per cent (19%) of the total number marking items Right in the fall marked the item Wrong or Omitted it in the spring.

There is no simple solution to this dilemma, but several actionable approaches relating to the scanning of the data for other evidence of a guessing tendency on the part of individual pupils may yet become clear as we proceed.

The inconclusive nature of the data that we are able to present here, while very helpful because it does reveal several lacks in the test and/or this experimental setup, simply tells us that there are too many uncontrolled factors to draw firm generalizations from such survey test results over short periods of time and without specific item selection to create a subset of items of unquestionable curriculum validity at the local level.

Table II-5

Analysis of Categories "WR+OR"* and "RW+RO"**

RANDOM SAMPLE

Test	No. of Items	Possible Pupil-Item Responses	Selected Pupil-Item Responses "WR+OR" No.	Mean	%	"RW+RO" No.	Mean	%
Word Meaning	38	21,546	5053	8.9	23	1694	3.0	8
Paragraph Meaning	60	34,020	7998	14.1	24	3590	6.3	11
Arithmetic Computation	39	22,113	5630	9.9	25	1781	3.1	8
Arithmetic Concepts	32	18,144	4059	7.1	22	2162	3.8	12
Arithmetic Applications	33	18,711	4158	7.3	22	2270	4.0	12

* Wrong or Omit fall, Right spring = possible gain

** Right fall, Wrong or Omit spring = possible fall guessing

For example, we do not know specifically the amount of time assigned to arithmetic instruction and we do not know to what extent other variables - such as the textbook, the general philosophy of the authors of these texts (traditional versus modern), or the competency of the teachers themselves - enter to determine the experimental results.

Some of these factors can not, and perhaps should not, be controlled for all children tested, but at least conditioning factors should be recognized.

Summary of Category Analysis

Each experimental evaluation of any Title I project (or similar local evaluation), as contrasted to comparison with a national norm, should be based upon a clear-cut statement of the objectives to be learned within the grade - while at the same time recognizing the fallacy of assuming that all children in the grade are equally capable of learning.

Tables similar to the three involved here, representing the performance of the random sample, are presented in Part III for the Title I group, and notable differences in the performance of the two groups will be evident at that time and can be discussed on their merits.

As expected, the Title I group performance is lower, testwise, but there are rays of hope in what appears to be improved learning in relation to known learning potential.

SOME CHARACTERISTICS OF STANDARDIZED TESTS RELEVANT TO THE RELATIONSHIP BETWEEN RIGHTS AND ATTEMPTS

Perhaps this section should be initiated by pointing out that the ideal relationship between rights and attempts, in the case of a standardized test, is largely a matter of attitude; attitude of the school administration, of the instructional staff, and of the pupils.

First of all, the purpose of any in-school test is to find out how much an individual knows about the body of information assessed by the test. This applies regardless of whether the test is a standardized test or is a local teacher-made test. Standardized tests, however, are constructed in such a way that certain factors are introduced which relate to, and affect the relationship between, rights and attempts; specifically, the almost universal use of some form of multiple choice test most of the time.

The very careful analysis, in the case of achievement tests, of the curriculum for the grade or grades in question prevents the introduction of material that is not pertinent to the universe of students to which the test is to be given. The test is often broken down into batteries covering one or two or, very rarely, three grades - each battery containing materials specifically identified with the instruction in that (those) grade(s).

It is legitimate to cover two or three grades in some subject tests, especially at the upper grades, because the curriculum sources from which the materials are collected are not specific enough to permit the assignment of a particular question or item to a particular grade in every instance. In such tests, the number of items should be greater than in other tests where there is more agreement as to grade placement.

The relevant fact here is that nothing ever gets into the preliminary experimental standardized test until it has been justified by determining that it does, indeed, appear in the appropriate curriculum material for that grade (or grades). Not just one or two textbooks are analyzed, but a large number of series are studied - together with courses of study and other relevant curriculum materials, including yearbooks of national societies and the like.

In fact the experimental editions, from the point of view of their comprehensiveness, may even be more curricula valid than the final editions of the tests, which are necessarily curtailed somewhat - due in part to the performance of the items when they are actually tried out in school situations, but also due to limits of length relative to other tests in the battery, time limits, and cost.

The aforementioned experimental editions for item tryout purposes require the arrangement of items in judged order of difficulty, so that the pupils taking the test do not find the items in random sequence. This is also a plus for the professional practices.

Subsequent to the item analysis and the re-examination of the items, those items or questions finally retained are arranged in a more precise, data based, order of difficulty - so that, ideally, except for the variations that exist from community to community, a child will answer, first, a very easy item, next, an item of somewhat greater difficulty, and so on, until he reaches the very hard items at the end of the test.

It is also customary to conduct experiments to determine the overlapping of scores of tests which are adjacent in a series. If the test is a comprehensive one, both as to variety of subject matter and range of grades covered, it is called a battery.

In the case of the Stanford Achievement Test, in general, each subject in each battery was administered to adjacent grades.

In the earlier days of testing (more than at the present, perhaps) a further experiment was carried out to determine the needed amount of time to answer the questions in each test - so that a statement like the following is commonly made: "In light of the fact that the items are arranged in order of difficulty, the time limit is always long enough so that a given child can answer correctly any items in the test which he is likely to know."

It is never considered desirable, from a test-maker's point of view, that the test score shall be enhanced by the effect of chance - although it is believed by this writer at this time, in terms of the data revealed by the present analysis, that altogether too much of this is taking place, an intolerable amount in point of fact.

The Rationale of Rights versus Attempts

If the points raised in the previous paragraphs are true as applied to a particular test, it seems quite evident that the important thing to determine for a test is how much time an individual needs to do all the items he is capable of doing. It is a good thing, rather than otherwise, to stop him before he has time to go on and guess on items of which he has no prior knowledge.

However, good or not, differentiated timing for individuals is something that is impossible to do - since the working time of individual pupils will vary so much from test to test or area to area. Giving unlimited time can disrupt a class because some (one or two) students per class dilly-dally along or are unable to complete a particular subtest other than by guessing, while others can consume enormous amounts of time.

It therefore follows, by logic alone, that if an individual answers Question #1 correctly, Question #2 correctly, Question #3 correctly, etc., until he has reached the point where he no longer knows the correct answer to most of the questions (and thus finds that he is beyond his depth either by knowing or reasoning) and then stops, the correlation (degree of "togetherness" or correspondence) between rights and attempts will be high.

Actually, how high the correlation will be will depend upon the temperament of the individual pupils and their willingness to recognize that they no longer are answering the questions on the basis of knowledge but are guessing randomly.

One would estimate, therefore, that the correlation between number right and number of attempted items in a valid test must be substantial; i.e., in the order of .85 or .90.

The Correction for Guessing

At this point we must interrupt this sequence of discussion to point out that for a period of years it was felt that a correction for guessing, such as rights-minus-some-fraction-of-the-wrongs, would counteract the occasional incident in which an individual would guess wildly instead of answering those items he knew and omitting the rest. 1/

Although the correction for guessing was largely dropped, generally nothing is said in the Directions to emphasize that guessing is not advisable or, in fact, is specifically mandated as being inadvisable. Certainly this was true of Stanford: Int. I: Form X. This is a great error in tactics, as will be seen as it is discussed later.

The Correlation of Rights versus Attempts for the Stanford Achievement Test: Intermediate I Battery: Form X: Grade 4

The author decided to determine, as the natural first step more than for any other reason, what the correlation between rights and attempts really was in this instance. He anticipated that the expected rather high correlations would result.

In order to do this task, since computer time was not immediately available, it was decided to use a population of 100 cases precisely, drawn randomly by sex; i.e., 50 boys and 50 girls. (The rosters were so ordered.) This sample was drawn and the correlations were worked out for the five tests with which the report is intimately concerned.

The resulting pattern of correlations (Table II-6) seemed to make no sense whatsoever. Even the highest of them fell far below the standard expected levels, and some of them were low enough as to make it not too unreasonable to ask if the correlations were significantly different from zero!

Even correlation ratios, unaffected by lack of normality and other population deviations, were computed without gaining any significant insight. The obvious negative skewness was not wholly overcome by the correlation ratios (there are two for each scatterplot).

It was felt that there must be something wrong with the sampling technique (although the writer could not discover any error) and, therefore, arrangements were made to re-do this part of the project by computer so as to involve the entire population instead of a sample of only 100 cases.

This set of calculations was done separately for the two populations with which this study is concerned; namely, the random sample of the state as a whole tested fall and spring and also the Title I children, similarly tested both fall and spring.

Table II-7 gives the results of the random sample analysis. It is perfectly evident that the second analysis strongly corroborates the analysis done the first time with respect to the low correlation values found.

There is a clear-cut difference in the r's for the last two math tests (namely Concepts and Applications) as compared to Word Meaning, Paragraph Meaning, and Arithmetic Computation.

Since the second set of correlations

1/ See Part I, pages 1 and 2.

Table II-6

Attempts versus Rights - 100 Case Random Sample

Correlations, Means, and Standard Deviations
RANDOM SAMPLE of 50 Girls and 50 Boys

Test	Corr. r	Mean		Standard Dev.	
		Attempts	Rights	Attempts	Rights
FALL:					
Word Meaning	.58	27.52	16.24	7.81	7.03
Paragraph Meaning	.41	45.98	24.82	11.53	9.28
Arithmetic Computation	.32	28.37	11.57	9.21	5.05
Arithmetic Concepts	.15	28.69	13.21	4.28	5.18
Arithmetic Applications	.18	28.28	12.50	5.93	5.34
SPRING:					
Word Meaning	.50	33.34	22.42	5.87	7.41
Paragraph Meaning	.17	53.78	33.10	6.86	10.69
Arithmetic Computation	.30	31.97	18.05	7.24	7.38
Arithmetic Concepts	.24	30.37	17.23	2.26	6.31
Arithmetic Applications	.18	30.98	16.79	2.33	6.32

had been done without paying any particular attention to the shape of the separate score distributions, we went back to our data to examine this parameter to see if we could find any causative factors that would result in this peculiar set of results.

Bivariate distributions were available only for the sample of 100 cases, but a more thoughtful examination of this small sample now revealed a potential piling up of cases at the top of the distribution on the attempts variable.

This led to the distribution of attempts alone on a univariate scale, the results of which are shown in Table II-8 (Distribution of Attempts) for the random sample.

Analysis of the Univariate Score Distribution for Skewness

On this table (II-8) the piling up became painfully evident - with a very large but varying proportion of youngsters attempting all of the items. This table, however, was not revealing with respect to the number of those who attempted all items but who, in turn, made high scores.

This led, then, to the separate distribution of the scores for those children attempting all items. The amazement of this writer was very great to discover that these reported scores ranged almost as widely as the distribution of raw scores on the test for the total group. See Table II-9 (Distribution of Right Responses for the Attempted ALL Group). Remember: We are now considering only the state random sample; the Title I group will be discussed later.

There were some few individuals who attempted all the items because they really were able to answer almost all of them correctly. Thinking specifically of the vocabulary test (Word Meaning), which had 38 items, earned scores of 35, 36 and 37 were found among the individuals who attempted all items in the spring.

The distribution of right scores for those who attempted all items revealed the obvious; i.e., much guessing had taken place and this indeed had inflated the scores for many of these individuals - although 15% earned scores which fell below the chance

Table II-7

Attempts versus Rights - 567 Case Random Sample

Correlations, Means, and Standard Deviations
RANDOM SAMPLE of 282 Girls and 285 Boys

Test	Corr. r	k*	Mean Attempts	Mean Rights	Standard Dev. Attempts	Standard Dev. Rights
FALL:						
Word Meaning	.55	(.84)	26.79	15.69	7.70	7.10
Paragraph Meaning	.49	(.87)	45.68	24.29	12.56	9.54
Arithmetic Computation	.29	(.96)	27.81	11.46	9.14	4.51
Arithmetic Concepts	.26	(.96)	29.26	12.91	4.75	5.20
Arithmetic Applications	.29	(.96)	29.13	12.77	6.19	5.12
SPRING:						
Word Meaning	.58	(.81)	33.34	21.78	6.34	7.34
Paragraph Meaning	.36	(.93)	54.03	31.90	9.79	10.53
Arithmetic Computation	.35	(.94)	32.08	18.25	7.76	7.13
Arithmetic Concepts	.33	(.94)	30.72	16.26	4.05	6.25
Arithmetic Applications	.25	(.97)	31.46	16.10	4.10	6.34

* Coefficient of Alienation

level (9.5) on the Word Meaning Test in the fall.

In other words, even though the scores were so low that they could have been reasonably gained by marking the answer sheet without regard to the test booklet, these students marked all items.

A Capsule Review of the Above

What we have now determined is that what was considered to be the normal pattern of rights versus attempts does not exist for the random sample population. For those children who mark-all-of-the-test-questions, the range of scores is almost as wide as the range for the total population - including the individuals who did not attempt all items!

The inevitable conclusion that must be drawn is that guessing was rampant in this population and that the general psychology for taking the test was one of: (1) answering an item without careful consideration of all alternatives if the answer was known; (2) if it was not known, then either estimation or sheer guessing was resorted to as a way of enhancing the individual's score on the test.

Let it be made abundantly clear that the fact that an individual marks every question on the test does not necessarily guarantee that he is a guesser as compared to one who attempts only items he reasonably thinks he can answer. This means our "attempt all" distributions are affected by the performance of the very able.

Let it be equally clear that if an individual marks all 38 items (on a test such as Word Meaning) and comes up with a final score that is at or near the guessing level or not far above it, the conclusion is equally inescapable that this result can come about only by a very inordinate amount of guessing.

Table II-8

Distribution of Attempts with Means and Standard Deviations

Random Sample of 567 Boys and Girls

No. of ATTEMPTS	Word Meaning F	Word Meaning S	Paragraph Meaning F	Paragraph Meaning S	Arithmetic Computation F	Arithmetic Computation S	Arithmetic Concepts F	Arithmetic Concepts S	Arithmetic Applications F	Arithmetic Applications S
60			132	279						
59			15	17						
58			19	20						
57			6	9						
56			12	11						
55			4	15						
54			3	11						
53			7	9						
52			3	4						
51			14	8						
50			12	13						
49			11	11						
48			33	29						
47			10	8						
46			11	12						
45			12	7						
44			16	7						
43			9	6						
42			13	6						
41			24	17						
40			18	6						
39			7	4	131	207				
38	124	269	20	10	29	27				
37	8	20	9	5	20	12				
36	4	27	5	4	7	19				
35	20	22	21	12	10	24				
34	17	17	8	1	10	17				
33	13	18	17	5	9	16			310	419
32	12	14	7	1	15	19	358	445	35	38
31	24	21	13	6	11	15	25	31	15	17
30	17	17	9	2	9	16	20	18	20	16
29	29	21	9	3	12	25	8	15	23	9
28	16	20	9		11	19	15	7	15	12
27	31	18	3		15	17	21	7	11	6
26	17	12	4		12	11	4	10	16	8
25	30	14	6		17	17	18	4	11	9
24	21	4	5	2	15	17	18	3	16	2
23	17	8	12	2	22	12	18	6	4	7
22	15	8	5	1	26	18	14	1	17	2
21	20	10	1		28	7	11	5	13	5
20	21	6	1	3	26	11	6	5	7	2
19	20	4	1		34	10	10		11	5
18	19	3	1		26	5	4	1	4	2
17	14	4	3		15	4	3		5	
16	13	2	1	1	12	6	2		7	2
15	4	2	1		14	2	1	1	4	
14	14	1	1		8	1	5		6	1
13	3				7	5	2	1	4	1
12	7		1		5				3	
11	6	1			4	2		1		
2-10	9	1			6	3	3	1	6	
N	565	564	564	567	566	564	566	562	563	563
Mean Att.	27.8	33.9	46.0	54.1	27.0	33.1	29.1	30.9	29.1	31.5
Std. Dev.	8.5	6.4	12.8	10.1	9.2	7.7	4.8	3.0	6.0	3.4
% Att.All	22%	48%	23%	49%	23%	38%	63%	79%	55%	74%

Table II-9

Distribution of Right Responses for Students Who Attempted All Items

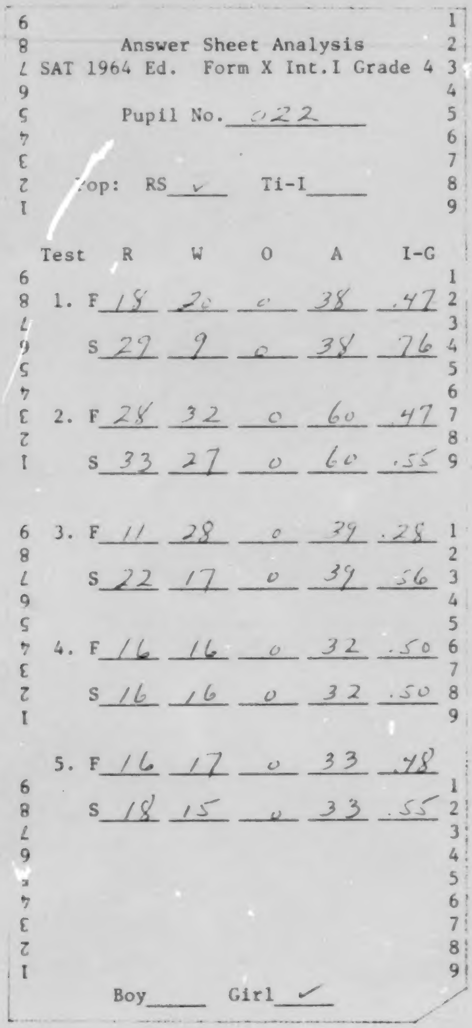
Random Sample of 567 Boys and Girls

No. of RIGHTS	Word Meaning (38)		Paragraph Meaning (60)		Arithmetic Computation (39)		Arithmetic Concepts (32)		Arithmetic Applications (33)	
	F	S	F	S	F	S	F	S	F	S
57				1						
56				1						
55				1						
54				2						
53				6						
52			1	1						
51				5						
50				7						
49			2	6						
48				6						
47				6						
46				7						
45			2	7						
44			3	7						
43			3	7						
42			6	14						
41			2	9						
40			1	8						
39			1	8						
38			2	10		1				
37		4	2	10		1				
36		2	5	10		1				
35	2	9	3	6		1				
34	1	10	1	3		1				
33	1	12	1	7		5				
32	1	13	4	7		5				1
31	3	19	6	5		2		1	1	1
30	3	6	3	10		8	1	5		4
29	4	21	7	11	1	4	1	9		4
28	10	16	7	2		4		6		3
27	7	16	2	7		5	2	5	1	10
26	5	18	3	8	1	9	4	16	1	14
25	7	14	7	10	1	5	1	16	2	11
24	3	13	3	4		8	7	19	4	16
23	5	11	5	9	3	12	2	15	7	23
22	2	8	4	3	2	3	13	16	4	26
21	1	10	4	6	1	12	12	29	8	19
20	6	14	6	4	1	10	9	21	13	16
19	5	7	8	9	5	9	15	26	11	21
18	5	3	9	6	6	14	27	25	12	24
17	6	8	3	2	4	6	15	30	29	18
16	5	7	5	5	4	8	14	26	13	24
15	3	6	5	5	6	9	15	21	20	23
14	6	7	3	2	9	12	17	11	27	13
13	6	4	2	2	11	11	30	20	11	21
12	5	4	1	2	17	8	32	27	20	15
11	1	2		2	12	14	19	18	29	13
10	5	2		1	12	4	26	21	13	26
9	5	1		1	12	2	25	24	27	22
8	7	1		1	7	3	20	11	20	18
7	3				5	3	22	13	12	12
2-6	1	1			11	7	29	14	25	21
Total N	124	269	132	279	131	207	358	445	310	419
%ofR.S.	22%	48%	23%	49%	23%	37%	63%	79%	55%	74%
Mean	19.6	25.1	27.1	33.9	13.4	19.3	13.5	16.9	13.4	16.5
S.D.	7.7	6.7	9.6	11.4	4.7	7.6	5.5	6.2	5.2	6.2

THE GUESSING INDEX OR THE RIGHTS/ATTEMPTS RATIO

Our studies to this point seemed to indicate the need for further investigation of the significance of the Rights versus Attempts information. Consequently, a new line of investigation was started; namely, a study of the behavior of a ratio comprised of the Rights divided by the Rights plus Wrongs, or Rights/Attempts (R/A).

In order to do this as expeditiously as possible without getting involved in machine analysis, the reverse side of some blank IBM cards were used to make up a record card for each pupil, a copy of which is shown in Figure II-5.



Answer Sheet Analysis
SAT 1964 Ed. Form X Int.I Grade 4

Pupil No. 022

Pop: RS ✓ Ti-I

Test		R	W	O	A	I-G
1.	F	18	20	0	38	.47
	S	29	9	0	38	.76
2.	F	28	32	0	60	.47
	S	33	27	0	60	.55
3.	F	11	28	0	39	.28
	S	22	17	0	39	.56
4.	F	16	16	0	32	.50
	S	16	16	0	32	.50
5.	F	16	17	0	33	.48
	S	18	15	0	33	.55

Boy Girl ✓

Figure II-5

Pupil Record Card

As shown, the card now contains the marginal information from the rosters - which consisted of Rights, Wrongs, and Omits - and from this we derived the number of Attempts by adding Rights and Wrongs and computed the R/A ratio by dividing the Rights by the Attempts.

This information was recorded on the cards, which were then sorted in R/A order separately by test. The distributions were transferred to Normal Percentile Charts and cumulative percents were calculated and plotted.

The plots were drawn from point to point to check on the shape of the distributions. Since this was done separately for fall and spring, each chart contained two distributions.

It turned out quite clearly that the amount of guessing involved, indicated by the diminution in the size of the R/A ratio as related to score, was more substantial in the fall than it was in the spring. This meant more guessing in the fall than in the spring.

Although one can rationalize as to the reason for this, this writer knows of no statistical method to arrive at any final explanation. Possibly the children, having a long summer vacation and being faced with content mostly related to the year ahead, were impelled to guess more in the fall in an attempt to make a good record.

One pair of distributions, plotted on a Normal Percentile Chart in the manner indicated, is reproduced as Chart II-1. Since all of the charts essentially follow the same general pattern, the others will not be reproduced.

It will be seen that the lines from the 10th to the 90th percentile ranks are fairly straight. (A straight line means a normal curve on these charts.) Any tendency to curvature in the line appears above or below these points.

This was also true of the charts which are not reproduced, so we can say essentially that the metric involved here, whatever it is, is one which is fairly symmetrically distributed. In other words, there is no skewness in the R/A ratios to reflect the skewness which was discovered when we made distributions of the "attempt" scores earlier.

The range of the R/A ratio is almost unbelievable. It goes from practically .00 to 1.00, indicating that the amount of guessing just has to be very, very substan-

Answer Sheet Study - II

Chart II-1

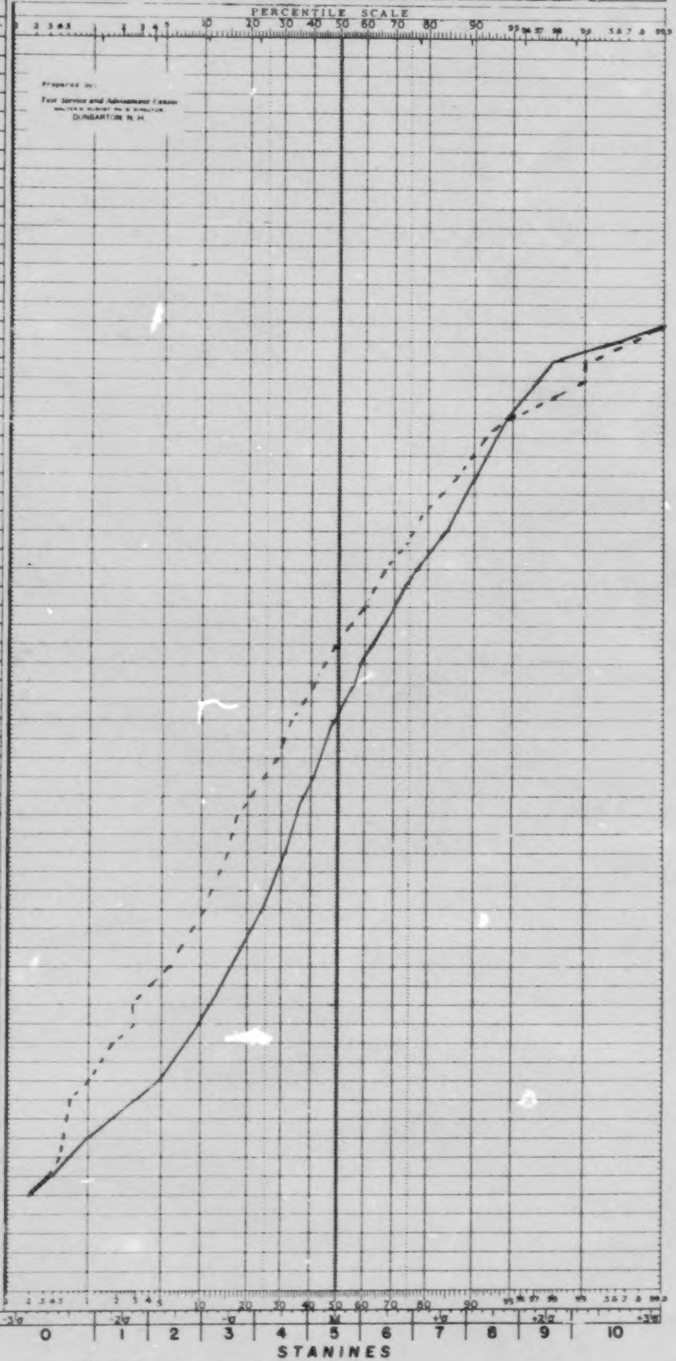
Plot of the Index of Guessing
Showing Decrease in the Amount of Guessing in the Spring Compared to Fall

RANDOM SAMPLE

NORMAL PERCENTILE CHART

	STUDY	Grade or Group	No. of Cases	MEASURE	Form	DATE	Graphs by	SCHOOL	COMMUNITY
I	Rights/Attempts-Word Meaning	4	565	SAT: Int. I	X	Fall '69	TS&AC	N.H. Statewide	Random Sample
II	" " - " "	"	"	" " "	"	Spr. '70	"	" "	" "

SCORE INTERVALS*					— I Fall			– – II Spr.		
			*		Freq.	Per-cent	Cum. %	Freq.	Per-cent	Cum. %
650-	325-	195-	130-	65						
640-	320-	192-	128-	64						
630-	315-	189-	126-	63						
620-	310-	186-	124-	62						
610-	305-	183-	122-	61						
600-	300-	180-	120-	60						
590-	295-	177-	118-	59						
580-	290-	174-	116-	58						
570-	285-	171-	114-	57						
560-	280-	168-	112-	56						
550-	275-	165-	110-	55						
540-	270-	162-	108-	54						
530-	265-	159-	106-	53						
520-	260-	156-	104-	52						
510-	255-	153-	102-	51						
500-	250-	150-	1.00-	50	9		100	1		100
490-	245-	147-	.98-	49	0		-	0		-
480-	240-	144-	.96-	48	8		98	7		99
470-	235-	141-	.94-	47	5		97	6		99
460-	230-	138-	.92-	46	10		96	9		98
450-	225-	135-	.90-	45	6		94	6		96
440-	220-	132-	.88-	44	10		93	19		95
430-	215-	129-	.86-	43	10		92	23		92
420-	210-	126-	.84-	42	11		90	23		87
410-	205-	123-	.82-	41	8		88	31		83
400-	200-	120-	.80-	40	14		86	12		78
390-	195-	117-	.78-	39	14		84	14		76
380-	190-	114-	.76-	38	18		81	35		73
370-	185-	111-	.74-	37	30		78	23		67
360-	180-	108-	.72-	36	17		73	18		63
350-	175-	105-	.70-	35	21		70	29		60
340-	170-	102-	.68-	34	25		66	29		55
330-	165-	99-	.66-	33	22		62	26		50
320-	160-	96-	.64-	32	17		58	16		45
310-	155-	93-	.62-	31	20		55	24		42
300-	150-	90-	.60-	30	19		51	25		38
290-	145-	87-	.58-	29	10		48	15		33
280-	140-	84-	.56-	28	18		46	12		31
270-	135-	81-	.54-	27	13		43	19		29
260-	130-	78-	.52-	26	20		41	22		25
250-	125-	75-	.50	25	15		37	19		21
240-	120-	72-	.48-	24	6		35	4		18
230-	115-	69-	.46-	23	15		33	7		17
220-	110-	66-	.44-	22	10		31	12		16
210-	105-	63-	.42-	21	19		29	11		14
200-	100-	60-	.40-	20	9		26	5		12
190-	95-	57-	.38-	19	15		24	12		11
180-	90-	54-	.36-	18	20		21	10		9
170-	85-	51-	.34-	17	13		18	10		7
160-	80-	48-	.32-	16	11		16	8		6
150-	75-	45-	.30-	15	11		14	5		4
140-	70-	42-	.28-	14	9		12	4		3
130-	65-	39-	.26-	13	11		10	1		3
120-	60-	36-	.24-	12	14		8	6		2
110-	55-	33-	.22-	11	6		6	0		-
100-	50-	30-	.20-	10	11		5	3		1
90-	45-	27-	.18-	9	6		3	1		.7
80-	40-	24-	.16-	8	2		2	0		-
70-	35-	21-	.14-	7	3		1	0		-
60-	30-	18-	.12-	6	2		.7	1		.5
50-	25-	15-	.10-	5	1		.4	1		.4
40-	20-	12-	.08-	4	1		.2	1		.2
30-	15-	9-	.06-	3	0		-	0		-
20-	10-	6-	.04-	2	0		-	0		-
10-	5-	3-	.02-	1	0		-	0		-
0-	0-	0-	.00-	0	0		-	0		-
MEDIANS					.60			.67		



*Graphs are plotted on the step interval checked at the top of the chart.

tial unless the logic which preceded this phase of our study - namely, that Rights and Attempts should not differ too greatly and that the correlations between the two should be essentially high, as they were not - was incorrect.

Means and standard deviations were also computed for each of the major groups, and these are shown in Table II-10.

It was clear from the above that the Rights/Attempts ratio had certainly earned a place in our consideration of the data involved in interpreting such test scores. Our first hope, that this would prove an effective substitute for the typical correction for guessing, proved to be a vain hope.

In the first place, high R/A ratios could be obtained by an individual who attempted a very small number of items but answered most of these correctly. This probably is a valid indication of this individual's tendency not to guess, especially considering the fact that the items are arranged in order of difficulty.

However, other instances where only a few items were attempted and only half, perhaps, of these were answered correctly indicated that the ratio certainly was not comparable from one part of the range to another, since a ratio of .50 based on 3 Rights and 6 Attempts is hardly a dependable statistic as compared to one based upon substantially larger numbers of Rights.

The ratio works best for the middle two-thirds of the distribution of Rights, or approximately plus and minus one standard deviation in the Rights or score distribution. The middle three stanines (roughly 54% in a reasonably symmetrical distribution) is also another way of selecting the place where R/A is at its best.

It is worth taking note of, however, for anyone who made an appreciable score above the average chance score, especially if the number of Wrongs is large. It certainly does indicate a temperamental tendency toward or away from random marking.1/

We cannot leave this matter without considering what the results obtained signify in terms of test-taking behavior. It would seem obvious from the above that we must build into the Directions for Administering a strict admonition not to respond on a purely guessing basis, since this randomness in the score distribution will only result in diluting the correlations between the before-after scores.

1/ It will be further noted that the R/A ratio equals the item difficulty when all items are attempted. A moment's thought makes this perfectly logical. In other words, to obtain item difficulties Rights are divided by all the total possible scores, which in effect is what happens when there are no omits.

Table II-10

Means and Standard Deviations - Rights/Attempts

RANDOM SAMPLE

Test	Fall		Spring	
	Mean	S.D.	Mean	S.D.
Word Meaning	.58	.213	.65	.181
Paragraph Meaning	.54	.174	.61	.181
Arithmetic Computation	.45	.192	.59	.207
Arithmetic Concepts	.45	.173	.53	.190
Arithmetic Applications	.45	.178	.51	.195

(It may be one of the major reasons why we got such peculiar correlations when we originally attempted to correlate Rights versus Attempts, as reported earlier.)

This applies even if the tests administered are taken a full year apart and are both different forms and different levels. Random guessing always reduces the correlation between pairs of scores.

Transforming the R/A Ratios Into Stanines

Although the distributions of scores show that these ratios generally are symmetrical and more or less bell-shaped, they certainly are not directly comparable to any of the other data that we have.

Since we do have stanines for most of the other data - e.g., Rights, Wrongs [1], etc. - that might be of value to the teacher, we used the data from the cumulative percentages on the Normal Percentile Charts to lay off stanine values for R/A and read off the stanine ranges.

Univariate distributions of stanines, showing them also graphically by means of asterisks, are shown in Figure II-6 for the fall and in Figure II-7 for the spring.

We need only to call your attention to the fact that the stanines did, indeed, fit the distributions remarkably well. (Compare theoretical with obtained frequencies.) The resulting stanine distributions are symmetrical, as of course they should be for stanines - which are, after all, normalized standard scores.

If the teacher has occasion to profile pupil results, these stanines are entirely appropriate for profiling purposes against any other set of data expressed in stanine form and based on essentially the same population.

For correlation purposes, however, they are of less value - since the R/A ratio has a built-in correlation with Rights and Wrongs because the denominator of the ratio is the sum of these two. Intercorrelations of subjects might work out well.

When Is a Test Invalidated By Guessing?

How high must the R/A score be to justify considering the test invalid? The only real way to resolve this problem is to examine the complete profile of the child, including a visual evaluation of his scores as listed on the roster.

In instances where we see a rather substantial run of Rights at the beginning of the test, we can more or less conclude that the child knew those particular items. This pattern of Rights will gradually break down as the items become harder, or in some cases it will suddenly break down, and the child either goes into a full guessing pattern or stops.

Perhaps we should solve this dilemma by considering "no test" any instance where a child has a Rights/Attempts ratio of .50 or less. A really satisfactory ratio should be .75 or higher, but apparently both teachers and pupils need much more understanding before such a high standard can be implemented.

(A high ratio means a large proportion of items Attempted were answered correctly.)

Children who are obviously guessing should be excluded from the item analysis and the N reduced by 1 for every such case eliminated in computing the difficulty values.

Any formula that can be devised to allow for guessing will work a hardship for some individuals.

For example, we have advocated the general thesis that a test should be relatively difficult at the beginning of the period of instruction in order to allow for plenty of room for the individual to indicate a real gain during the period of time he is subsequently under instruction, provided (1) that he is subsequently exposed to instruction, and (2) he is believed to have reached a level of mental development to permit learning the content in question.

---

[1] Available on request.

Answer Sheet Study - II

Sta-nine	R/A Range	Frequency		Percent		
		Act.	Theor.	Act.	Theor.	
WORD MEANING:						
9	.94-1.00	22	23	4	4	*****
8	.86-.93	36	37	6	7	*********
7	.75-.85	75	68	13	12	*******************
6	.67-.74	94	99	17	17	***********************
5	.54-.66	108	112	19	20	***************************
4	.39-.53	97	99	17	17	************************
3	.29-.38	72	68	13	12	******************
2	.22-.28	35	37	6	7	*********
1	.08-.21	26	23	5	4	******
Median=.60		565				
PARAGRAPH MEANING:						
9	.85-.96	23	23	4	4	******
8	.77-.84	36	37	6	7	*********
7	.70-.76	74	68	13	12	******************
6	.62-.69	93	99	16	17	***********************
5	.50-.61	109	111	19	20	***************************
4	.40-.49	104	99	18	17	**************************
3	.31-.39	65	68	12	12	****************
2	.25-.30	37	37	7	7	*********
1	.07-.24	23	23	4	4	******
Median=.54		564				
ARITHMETIC COMPUTATION:						
9	.81-1.00	24	23	4	4	******
8	.72-.80	37	37	7	7	*********
7	.60-.71	68	68	12	12	*****************
6	.48-.59	98	99	17	17	************************
5	.37-.47	113	112	20	20	****************************
4	.28-.36	105	99	18	17	**************************
3	.23-.27	62	68	11	12	***************
2	.16-.22	37	37	7	7	*********
1	.06-.15	22	23	4	4	*****
Median=.42		566				
ARITHMETIC CONCEPTS:						
9	.76-1.00	22	23	4	4	*****
8	.68-.75	38	37	7	7	*********
7	.59-.67	69	68	12	12	*****************
6	.49-.58	100	99	13	17	*************************
5	.39-.48	107	112	19	20	***************************
4	.31-.38	100	99	18	17	*************************
3	.23-.30	66	68	11	12	****************
2	.19-.22	43	37	8	7	***********
1	.06-.18	21	23	4	4	*****
Median=.42		566				
ARITHMETIC APPLICATIONS:						
9	.77-1.00	25	23	4	4	******
8	.68-.76	34	37	6	7	********
7	.60-.67	71	68	13	12	******************
6	.49-.59	101	98	18	17	*************************
5	.39-.48	109	111	19	20	***************************
4	.30-.38	96	98	17	17	************************
3	.24-.29	69	68	12	12	*****************
2	.18-.23	36	37	6	7	*********
1	.06-.17	22	23	4	4	*****
Median=.44		563				

(Each * represents 4 students)

FIGURE II-6

Stanine Distributions - Index of Guessing or Rights/Attempts
RANDOM SAMPLE - Fall 1969

Answer Sheet Study - II

Sta-nine	R/A Range	Frequency		Percent		
		Act.	Theor.	Act.	Theor.	
WORD MEANING:						
9	.92-1.00	23	23	4	4	******
8	.87-.91	40	37	7	7	**********
7	.82-.86	62	68	11	12	***************
6	.72-.81	102	99	18	17	*************************
5	.63-.71	121	112	21	20	******************************
4	.53-.62	94	99	17	17	***********************
3	.40-.52	60	68	11	12	***************
2	.32-.39	40	37	7	7	**********
1	.09-.31	23	23	4	4	******
Median=.67		555				
PARAGRAPH MEANING:						
9	.88-.98	24	23	4	4	******
8	.83-.87	31	37	5	7	********
7	.77-.82	72	69	13	12	******************
6	.68-.76	101	99	18	17	*************************
5	.59-.67	111	112	20	20	****************************
4	.46-.58	97	99	17	17	************************
3	.35-.45	74	69	13	12	******************
2	.26-.34	37	37	6	7	*********
1	.13-.25	20	23	4	4	*****
Median=.62		567				
ARITHMETIC COMPUTATION:						
9	.91-.97	23	23	4	4	******
8	.86-.90	31	37	5	7	********
7	.77-.85	73	68	13	12	******************
6	.68-.76	94	99	17	17	***********************
5	.54-.67	124	111	22	20	*******************************
4	.41-.53	93	99	16	17	***********************
3	.30-.40	67	68	12	12	*****************
2	.23-.29	36	37	6	7	*********
1	.10-.22	23	23	4	4	******
Median=.60		564				
ARITHMETIC CONCEPTS:						
9	.85-.94	23	23	4	4	******
8	.78-.84	42	37	7	7	**********
7	.69-.77	65	68	12	12	****************
6	.58-.68	97	93	17	17	************************
5	.48-.57	105	111	19	20	**************************
4	.38-.47	103	98	18	17	**************************
3	.29-.37	60	68	11	12	***************
2	.21-.28	49	37	9	7	************
1	.13-.20	18	23	3	4	****
Median=.52		562				
ARITHMETIC APPLICATIONS:						
9	.83-.97	24	23	4	4	******
8	.77-.82	34	37	6	7	********
7	.68-.76	62	68	11	12	***************
6	.59-.67	91	98	16	17	***********************
5	.46-.58	122	111	22	20	******************************
4	.34-.45	100	93	18	17	*************************
3	.25-.33	70	68	12	12	*****************
2	.19-.24	36	37	6	7	*********
1	.12-.18	24	23	4	4	******
Median=.52		563				

(Each * represents 4 students)

FIGURE II-7

Stanine Distributions - Index of Guessing or Rights/Attempts
RANDOM SAMPLE - Spring 1970

USE OF PREDICTED SCORE TO DETERMINE EXTENT OF GUESSING

Some years ago (1950 approximately), this writer devised a technique for estimating an untimed score from a time-limited score - after observing that some children, particularly the slow learners, were handicapped because the time limits (normally quite satisfactory) were, for them, unduly short.

This technique is expressed in the formula: Untimed Score = A + (B/C x D), where

A = the score earned within the stated time limits to the beginning of the series of omitted (hard) items;
B = the score earned on the last twenty items attempted;
C = the sum of the percent passing the last twenty items attempted; and
D = the sum of the percent passing the remaining (i.e., not attempted) items.

A closer look at this formula makes it obvious that the value "C" is actually the mean score earned by the population in question on a subset of twenty items IF THESE ALWAYS ARE THE SAME TWENTY ITEMS - since the sum of the difficulty values of any group of items taken by a defined population, with the decimal point retained for each percent, is the mean score for that population. Similarly, "D" is the mean score of the items not attempted.

The precise effect of the application of this formula, under the condition stated, is to estimate a score for the items not attempted by saying that it would be some parameter of the mean score for the selected twenty items, dependent upon the "goodness" of the performance of the individual on the chosen subset of twenty items - as indicated by the ratio B/C.

Note: The subset of items used in the original application of the formula was the last twenty items attempted by each individual. Just what subset of items was used for Student A or B or C was irrelevant, provided that it gave a good estimate of the ability of the individual to answer the questions contained within the final omitted items. All twenty items were supposed to be "attempted" in the sense that the individual had carefully considered each. Omits in small numbers were permitted. If the "DK" (Don't Know) space was available, this could be used to a reasonable extent. The score earned, except for the minimized effects of chance, was probably the optimum estimate of the quality of work the individual was capable of doing on the test, allowing for unreliability and the failure of the individual to attempt all items.

While this formula had the virtue of correcting the individual's score so that it gave a reasonably close estimate of what he was capable of doing, it would underestimate in most instances the score earned by the very ablest individuals. This was true because difficulty values for "D" were easy for the ablest students but difficult for the average or low achievers. This was not a serious limitation, however, because these very able individuals almost always did all they were capable of doing in the time allowed and time, thus, was itself not a factor. 1/

In this present situation, the analysis of the performance of the individuals comprising the random sample from the total New Hampshire state population at the fourth grade level indicated that a very much larger proportion of individuals answered more of the items correctly than one would anticipate if guessing were not present as a common practice.

To put this differently, one would anticipate that both the rights score and the attempts score would be more or less normally distributed if the items were answered on the basis of true information or knowledge or on the basis of a rational analysis of the alternatives - with the final choice being made on the basis of some knowledge, if not total knowledge. The rights score and the attempts score would correlate highly. It has been shown that this was not the case for the New Hampshire random sample on any of the five tests analyzed.

In search of some additional light on this subject following the analysis of the distributions of scores for the "attempt all" population, the writer has adapted the formula described above for the estimation of a total score on a test - using as the basis for the estimate the first twenty items in each test (in Paragraph Meaning, 23), which constitute the easy items.

It is felt that guessing tendency would be minimized in answering this subset of items, because a much larger number of individuals would know the answers to a very substantial proportion of the items selected and would, therefore, not be likely to resort to guessing.

1/ While this procedure has not been published, it was first described in 1949-50 and a nomograph, plus cumulative sums of percents passing, facilitated the free choice of any subset of twenty items as indicated.

Note: Items were arranged in order of difficulty on the regularly published edition of Stanford: Intermediate I: Form X. If the difficulty values of these easy, or at least easier, items is used constantly as the basis for the estimate of total score, it is possible to arrive at an estimate for individuals of varying levels of ability and compare this with the score they actually earned.

In this procedure, the values in "A" and "B" are always based upon the first twenty items (23 for Paragraph Meaning so as to include all questions on the last paragraph in which the twentieth item occurred).

Thus, the original formula is modified to: Predicted Score = $A + (A'/C \times D)$, where

A = the score earned on the first twenty (or 23) items;
A'= the same value;
C = the mean score of this population on the first twenty (or 23) items;
D = the mean score of this population on the remaining items - e.g., the last 18 items in Word Meaning.

In making the decision to use the first twenty (or 23) items as a constant in this study, two factors were involved. First, guessing certainly would be minimized by using the very easiest items; secondly, the difficulty range of such items must yield enough variation of score to be reasonably reliable.

This estimation process, when done by hand, proved to be a time-consuming task. In this report for the random sample we will give the results for three tests only; namely, Word Meaning, Paragraph Meaning, and Arithmetic Computation.1/ Word Meaning and Paragraph Meaning together constitute Reading, the subject of greatest concern to Title I programs.

As earlier indicated, the special reason for concentrating on these subtests was the obvious difference in the relative environmental impact. Word Meaning and Paragraph Meaning are greatly affected by the total environment; Arithmetic Computation is almost exclusively school oriented.

It is well known that generally, in standardized tests, great care is taken in the final forms of a test to arrange the items in order of difficulty.

There are various reasons for this, some statistical and some psychological. It would be psychologically unwise, for example, to begin a test with an extremely difficult item; this would immediately discourage the child in taking the rest of the test. Even a chance arrangement of items with respect to difficulty would have much the same effect.

On the other hand, by arranging the items in order of difficulty (easiest first) every child would be encouraged to do what he is capable of doing. This arrangement has the additional advantage that it makes the time limit of much less importance, since in almost every instance a child will get all of the items right that he can honestly answer correctly in the time allowed, even if he does not attempt all of the items.

This has been shown repeatedly in experiments to determine the effective working time limits, which are so necessary for the practical administration of a test.

It is important to point out also that it is standard operating procedure to build into a standardized test a wide enough range of difficulty to provide for both the least able and the most able pupils within the group to be tested, obviously a necessity in a survey test.

Usually this is accomplished at the lower end of the scale by including some items that should have been learned at a grade previous to that at which the test is normally given.

Ideally, in any local before-after program all the items should be validated against the local curriculum; but this is rarely done, unfortunately.

The upper levels are provided for by making items of greater difficulty while still staying within the curriculum normally found within the grade or grades for which the test is intended; *i.e., to avoid as much as possible including any items to which the target group has not been exposed to instruction.*

Consider the difference in the difficulty of adding two three-place numbers and ten three-place numbers. In the first instance, the opportunities for error are fairly limited; while in the second, the opportunities for making errors are greatly increased because of the number of times in which an individual must perform the basic

---

1/ Work was completed on the remaining tests after this manuscript was completed, and the results are given in Appendix C.

operations involved in solving a problem of this nature.

Both types, however, involve exposure to the basic problem of complying with locally made "behavioral objectives" - which, far too often, are rather inadequately conceived.[1]

The analysis that follows has divided the total random sample into subpopulations of boys and girls, and each of these into those who Attempted-All items versus those who Did-NOT-Attempt-All items.

The selection of the Attempted-All group as a way of designating the guessing group is an impure or contaminated way of identifying guessers, because some who Attempted-All items did not guess in a random fashion at all, or did so on the very few items at the end of the test.

These latter are the very able children, who naturally earn high or near perfect scores because they know the answers or can arrive at them rationally by thinking about the alternatives offered and choosing one of two alternatives.

The technique of predicting a score and comparing the predicted score with the actually earned score works best to identify the guessers where the "attempted" count is substantially higher than the earned score. The greatest difference, obviously, is to be found where all items have been attempted and few are right. (This has been previously considered in discussing the R/A ratio.)

The correlations reported in this study have been arrived at by actually plotting the data on bivariate charts. Computerized calculation would have been much faster and possibly more accurate.

Correlation coefficients alone can be very misleading, especially the coefficient without the corresponding plot. For example, a correlation coefficient does not as a general rule reflect gain in score but simply expresses rank order.

Simplified computational formulas were chosen to make it possible to obtain the correlations manually from the plotted charts with a minimum of work and to check the r's by the use of several different formulas, all derived from sums and differences of scores.

This computational process also yielded means and standard deviations, which are helpful in studying changes in magnitude and variability of scores.

The correlations reported are listed in Table II-11, so that someone can see at a glance the rather substantial number of populations separately studied and the general trend in the r's for different subsamples.

The number of cases, the means, and the standard deviations are reported separately in Table II-12, immediately following.

No attempt is made to evaluate statistically the differences between the means of the Attempted-All group and the Did-NOT-Attempt-All group, because we are not dealing with purely random samples and we had no reason to anticipate, without investigation, that the distributions on which the correlations were based even were normal or similarly skewed, so as to provide a rectilinear plot.

Considering first the correlations alone (Table II-11), it is noteworthy that they are high in practically every comparison; as a matter of fact they are surprisingly high, all things considered.

One might even conclude that the first twenty items on a test give about as good a measure as the total test, at least for the tests considered.[2] This would not, of course, be true - because such a procedure does not take account of the range of performing ability on the whole test for the group from which the New Hampshire item difficulties were derived; i.e., the random sample of pupils tested at grade 4 in 1969-70.

If we were to consider the distribution of scores for the first twenty items only, we'd find many of the ablest children piling up at the top score of 20 and our predicted score would be (and is) too low. The predicting formula helps, but it still fails to do justice to the very ablest children - whose predicted scores regularly fall below their earned scores.

---

[1] This may be conceived as a criticism of the local "curriculum objectives committee" (by whatever name), but it is not so intended. Realistic "behavioral objectives" are time-consuming and difficult to prepare, especially if one keeps in mind the subsequent need to evaluate success or failure. Nebulous objects defy evaluation!

[2] In some cases they approach or exceed reported reliability coefficients.

Table II-11

Correlation Coefficients

Actual Scores versus Predicted Scores

RANDOM SAMPLE

	Attempted-All		Did-NOT-Att.-All		Total Group	
	Fall	Spring	Fall	Spring	Fall	Spring
Word Meaning						
Boys	.94	.92	.93	.89	.92	.89
Girls	.95	.88	.92	.90	.90	.89
Paragraph Meaning						
Boys	.90	.90	.81	.78	.79	.82
Girls	.90	.91	.80	.73	.77	.81
Arithmetic Computation						
Boys	.91	.92	.92	.89	.89	.87
Girls	.89	.89	.92	.89	.89	.88

Next it should be noted that the correlations for the Attempted-All and Did-NOT-Attempt-All groups combined (i.e., the Total Group) tend to drop slightly, but only a point or two in the hundredths place. Most readers would regard such small differences as being practically insignificant.

Absolute Changes in Score Over a Given Time

In order to establish the amount by which individuals change their status by gaining additional points of score over the intervening seven months, one must look at the data in the table of means and standard deviations (Table II-12), given separately to avoid clouding the issue of the level of agreement of actual versus predicted scores. These are very important data, however, and need careful study.

In Table II-12 we have summarized a very large amount of data in what would be called a general purpose table; that is, one that presents far more data than can be efficiently discussed in detail in the text. Thus it presents the reader with a challenge to search the table for meanings not specifically brought out in the discussion.

Most Significant Elements in Table II-12

The table contains data relevant to the actual or recorded score (that is, number right) - first, for those children who Attempted-All items; secondly, for those who Did-NOT-Attempt-All; and finally, for Total Group. These data are given separately for fall and spring as well as for boys and girls, together with numbers of cases in each subgroup.

We have then added to this table comparable data for the predicted scores as for actual or reported scores.

We have given, finally, the differences between means for both the actual raw score earned (i.e., the number right as scored by the machine) and the predicted score separately for the Attempted-All group versus those who Did-NOT-Attempt-All and the Attempted-All group versus the Total Group of boys or girls.

(Note: In the latter case, the Total Group data includes both of the previous subsamples; thus in a sense this column, then, is diluted by the inclusion of the Attempted-All group. In point of fact this final comparison does, however, indicate the effect of the inclusion of the Attempted-All subgroup data on the total results as previously reported to the community.

Thus it highlights the fact that these children do constitute a separate subpopulation, distinct in character from the

Table II-12

Comparison of Attempted-All versus Did-NOT-Attempt-All Groups for Selected Statistics With Particular Emphasis on Magnitude and Direction of Differences Between Means

RANDOM SAMPLE

	Attempted-All			Did-NOT-Att.-All			Diff.of	Total Group*			Diff.of
	N	Mean	S.D.	N	Mean	S.D.	Means1/	N	Mean	S.D.	Means2/
Word Meaning											
BOYS											
Fall - Actual Score	62	19.9	8.2	222	14.6	7.1	+5.3	284	15.8	7.7	+4.1
Predicted Score		16.8	6.8		15.5	6.6	+1.3		15.8	6.7	+1.3
Spring-Actual Score	136	25.2	7.2	148	18.5	6.5	+6.7	284	21.7	7.6	+3.5
Predicted Score		23.2	5.9		20.3	5.9	+2.9		21.7	6.1	+1.5
GIRLS											
Fall - Actual Score	62	19.4	7.2	219	15.0	6.1	+4.4	281	16.0	6.6	+3.4
Predicted Score		16.5	6.2		15.9	6.0	+ .6		16.1	6.0	+ .4
Spring-Actual Score	133	25.3	6.1	147	19.2	6.3	+6.1	280	22.1	6.9	+3.2
Predicted Score		23.7	4.7		20.5	5.6	+3.2		22.0	5.4	+1.7
Paragraph Meaning											
BOYS											
Fall - Actual Score	66	27.5	10.4	217	22.4	9.3	+5.1	283	23.6	9.8	+3.9
Predicted Score		22.3	8.5		24.0	8.2	-1.7		23.6	8.3	-1.3
Spring-Actual Score	138	33.3	11.8	147	29.4	9.5	+3.9	285	31.1	10.9	+2.2
Predicted Score		31.0	10.2		32.5	8.0	-1.5		31.8	9.1	- .8
GIRLS											
Fall - Actual Score	66	27.1	8.6	215	24.4	9.2	+2.7	281	25.1	9.1	+2.0
Predicted Score		21.9	7.5		25.9	6.9	-4.0		24.9	7.2	-3.0
Spring-Actual Score	141	34.6	10.7	141	30.5	8.8	+4.1	282	32.6	10.0	+2.0
Predicted Score		32.2	8.9		32.7	7.2	- .5		32.4	8.1	- .2
Arithmetic Computation											
BOYS											
Fall - Actual Score	72	12.4	4.9	213	10.8	4.1	+1.6	285	11.2	4.4	+1.2
Predicted Score		10.6	4.9		11.3	4.8	- .7		11.1	4.8	- .5
Spring-Actual Score	113	18.7	7.8	170	16.9	6.2	+1.8	283	17.7	6.9	+1.0
Predicted Score		16.2	6.5		17.9	5.7	-1.7		17.2	6.1	-1.0
GIRLS											
Fall - Actual Score	59	12.7	4.6	221	11.5	4.6	+1.2	280	11.8	4.6	+ .9
Predicted Score		10.9	4.7		12.2	4.5	-1.3		12.0	4.6	-1.1
Spring-Actual Score	94	20.1	6.9	187	18.6	6.9	+1.5	281	19.1	6.9	+1.0
Predicted Score		18.9	5.2		19.3	5.7	- .4		19.1	5.5	- .2

* Bivariate Overlays showing displacement of "Attempted-All" group versus "Did-NOT-Attempt-All" group shown in Appendix for Word Meaning and Arithmetic Comp.

1/ "Attempted-All" group versus "Did-NOT-Attempt-All" group

2/ "Attempted-All" group versus "Total Group"

rest, which unwittingly has affected the performance of the total group because of the very significant difference as to their method of marking the answer sheet.)

At this point, since we have considered the gains from fall to spring for the total random sample group elsewhere, let us concentrate mainly on the data for the Attempted-All versus the Did-NOT-Attempt-All groups.

(Only three of the five tests generally analyzed in this report are given in Table II-12. Arithmetic Concepts and Arithmetic Applications have been examined closely enough to see that their results are consistent with the others.)1/

First notice that for all subgroups on all tests the differences between means of actual scores favor the Attempted-All group. This is true even when we consider the Attempted-All group in comparison with the Total Group, of which they are a part. In Word Meaning, this is true of predicted scores as well.

When we move on to Paragraph Meaning and Arithmetic Computation, we see negative differences between means for all predicted scores are higher for the Did-NOT-Attempt-All group and the Total Group than for the Attempted-All group. Out of sixteen comparisons, all are negative.

Note that in the table the negative differences always apply to the predicted score, not the actual score; i.e., the number right. The significance of this is that early performance predicts lower total scores for the Attempted-All group than for those who Did-NOT-Attempt-All or the Total Group.

The question remains, however, (and must remain unanswered in this report) as to which of the sets of scores, actual versus predicted, is the more valid measure of group or individual performance.

The writer's guess is that the predicted score truly represents the performance of a child more adequately than actual score when his earned score (i.e., the number of items answered correctly as scored) is on the low side. High scores (i.e., 70% to 75% right) are an exception almost by definition on a standardized survey-type test. Otherwise, the ablest children would not be measured!

Reproduction of the bivariate charts for this report presented a very difficult problem. Separate Attempted-All versus Did-NOT-Attempt-All bivariates illustrate clearly the effect of guessing, but it was hard to compare two charts.

Sample Bivariate Charts

Actually, each bivariate chart as shown in Appendix B consists of two bivariates combined; one for those who Attempted-All of the items superimposed on the chart for those who Did-NOT-Attempt-All items.

The Attempted-All group is printed in a contrasting color so that one can see the change in the distribution from group to group, which always is in the direction of higher overall performance for those who Attempted-All items and, therefore, took advantage of every opportunity to guess. It is this spurious gain due to guessing which must be identified and eliminated to make the test truly valid.2/

Perhaps these bivariate distributions, from a layman's point of view, are the most significant or convincing evidence of the presence and effect of guessing.

It would have been very desirable to reproduce in the report the entire 96 bivariates from which the correlations were computed. This consists of 36 such charts for the random sample group, with which we are presently concerned, and 60 for Title I, which will be the concern of the next section. However, this was impractical from a space point of view and, therefore, only a selection of these have been reproduced.

The consistent offsetting of the Attempted-All subpopulation is conspicuous on these charts. Each bivariate group (i.e., black versus colored) taken by itself yields a correlation most of the time higher than for all cases combined. The raw score means of the Attempted-All group are higher than the raw score means of the Total Group.

The greatest spurious gains are for those who EARN low scores; thus, guessing hurts most those children who are in greatest need of help!

1/ As time permits, Arithmetic Concepts and Applications will be completed; but the three tests shown were enough to demonstrate the essential fact with respect to the uniqueness of the Attempted-All subgroup. Costs and time led to the decision to omit the remaining two tests for the moment from the random sample analysis.

2/ Bivariate charts are shown separately for boys and girls and for two subjects only, Word Meaning and Arithmetic Computation.

We must conclude, therefore, from these data taken as a whole, that guessing, certainly in the sense of marking every item in the test regardless of whether you know the answer or not, generally does have the effect of raising one's apparent score and, therefore, getting a higher percentile rank or grade equivalent.

Therefore guessing misleads the teacher as to what the individual really knows. When the score data are reduced to item analysis information, as they have been in this study, such contamination has a very significantly detrimental effect.

As we continue our inspection of the bivariate charts, we must note that all of the changes found from the actually earned score to the predicted score are not always in the direction of an increase in predicted score. This is in support of what we said earlier - that this is an impure or contaminated way of identifying guessing youngsters.

One of the factors responsible for a drop in the predicted score as compared to the actual earned score is that some of the items remaining after the first twenty items were scored were too hard for all but the most able children in the population. Thus the very able pupils actually did earn scores higher than their predicted scores (and some earned a nearly perfect score).

Most significant, however, is the fact that the intercorrelation between purely guessed scores is zero and will vary from this value only by chance. 1/ Partial guessing either in fall or spring reduces all correlations in a manner proportionate to the amount of guessing.

1/ Not demonstrated here. See page II-5.

EVALUATING PUPIL PERFORMANCE FROM THE PUPIL ROSTERS

The basic data for this study consisted of a listing of the response of each pupil to each item - not by rights, wrongs, or omits but actually by tabulating the number of the alternative chosen by each pupil (1, 2,3,4, or 5) - and having the tabulator insert the correct response for all items after every fifth pupil entry, so that it is possible to determine by reference to this key whether a child has answered any item correctly and, if not, which of the alternatives he has chosen.[1]

This type of listing is very essential for certain aspects of the analysis we have done, and in particular the analysis relating to the categorization of the items as Right in the fall, Right in the spring, etc.

In the original mode (Figure II-8), the chief advantage was that it identifies the Wrongs by all alternatives other than the correct response and indicates distractors that are working effectively versus those which appear not to be attractive to anyone except on the basis of pure chance.

It also has the great advantage to allow the person constructing a test to spot instances where the keyed response may not be correct or where there may be more than one correct response - since in such instances the number of children choosing a particular option may be out of line with what would be expected for the difficulty value of the item as a whole.

A different approach to reporting item analysis data that lists only Rights, Wrongs and Omits makes the examination of the pupils' responses much easier than the approach we have used here. In order to illustrate this, the same page of selected cases is shown in the Right, Wrong, Omit mode. (Figure II-9)

The item by item comparison of the performance of each child, fall and spring, constitutes the ultimate approach to the problem of evaluating the amount of guessing present.

It is evident from the rosters that as one moves from left to right across the page (that is, from the early items to the later ones), the number of Right responses definitely decreases and the number of Wrong or Omitted responses increases.

Choosing any particular case, it is quite evident that the proportion of the first twenty items which are answered correctly is much greater than the proportion of the remaining items in the test (13 in the case of Arithmetic Applications, for example). This follows from the publisher's arrangement of the items in order of difficulty.

We know the "average" guessing score from the number of alternatives in relation to the number of items in the test, assuming random marking.

In the Arithmetic Applications Test, for example, which is a five-choice multiple choice-type test with a total of 33 items, we would expect by chance on the average one out of each five items marked to be the correct response, also assuming random marking totally; i.e., none marked from sure or partial knowledge.

The average guessing score for a test of 33 five-choice items would be one-fifth of the number of items, or 6+.

One can say without equivocation that anyone who has a score of only 6, where the number of items answered correctly is scattered across the listing of items and not bunched at the beginning of the test, is surely guessing and the test should be considered to be invalid.

At the other extreme, if the first six items were answered correctly and very few additional responses were Wrong and most remaining items were Omitted, this would suggest poor performance but little or no guessing.

The illustration we have chosen to use in this particular section is taken from one page of the Arithmetic Applications pupil roster. This is a five-choice test where one of the responses always is "Not Given." Since "NG" is a scored response, however, it is considered the same as the other responses.

To recapitulate the obvious, counting the number of "R's" across the sheet to the right gives the total number of Right responses, or the individual's score. Similarly, counting the "W's" gives the number of responses that are incorrect, and counting the number of "O's" gives the number of items that were Omitted.

In a comparable fashion, counting the "R's" in a particular column for the total population gives the number of individuals answering the item correctly.

---

[1] See pages 9 and 10, Part I.

Answer Sheet Study - II

```
ITEM NUMBER 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33    TOTALS

 KEY        1 3 5 3 1 2 1 1 5 2  2  4  1  3  4  3  5  4  3  3  4  5  2  4  3  2  1  2  1  3  4  4  2    1   X   0
PUPIL F-S

 001  F     1 3 5 3 1 2 1 5 5 3  5  4  1  2  4  5  5  3  5  5  4  5  5  2  3  5  1  4  5  0  0  0  0   16  13   4
 001  S     1 3 5 3 1 2 1 1 5 2  2  4  1  1  4  3  5  4  2  3  4  5  2  5  5  1  3  4  1  5  1  4  3   23  10   0

 003  F     2 5 1 5 1 2 1 4 0 4  0  0  0  2  3  3  5  4  3  0  0  0  0  0  0  0  0  4  5  3  0  0  0    8  10  15
 003  S     1 3 5 5 1 3 1 4 3 5  2  3  3  5  4  3  5  4  3  3  4  5  3  2  2  3  3  4  3  3  3  2  4   15  18   0

 004  F     1 3 1 3 1 4 1 2 5 3  2  5  1  3  4  3  5  4  3  5  4  3  2  5  4  5  5  4  0  0  0  0  0   16  12   5
 004  S     1 3 5 3 1 5 1 1 1 2  2  4  1  3  4  3  5  4  3  3  1  5  2  1  1  5  1  2  5  3  4  1  1   24   9   0

 006  F     1 3 5 5 1 2 1 5 4 0  5  5  1  3  4  5  5  3  5  5  5  4  5  5  5  2  1  5  5  5  4  1  1   13  19   1
 006  S     1 3 5 3 1 5 1 1 1 2  2  4  1  3  4  5  5  3  5  3  [illegible]  5  2  1  1  5  1  2  5  3  4  1  2   22  11   0

 008  F     1 3 5 3 1 2 1 4 5 2  5  4  5  3  3  3  5  4  3  3  4  5  5  5  5  2  3  4  4  3  0  0  0   20  10   3
 008  S     1 3 5 4 1 3 1 1 1 2  2  4  1  3  3  3  2  5  0  4  4  2  2  1  1  5  1  2  5  3  4  3  2   19  13   1

** KEY **   1 3 5 3 1 2 1 1 5 2  2  4  1  3  4  3  5  4  3  3  4  5  2  4  3  2  1  2  1  3  4  4  2

 011  F     1 3 5 3 1 4 1 1 1 3  1  4  1  2  3  3  5  4  3  5  5  4  3  3  4  5  5  4  5  5  5  4  3   14  19   0
 011  S     1 3 5 4 5 4 1 4 5 2  5  4  5  3  4  3  5  4  3  2  5  5  2  1  5  1  5  4  5  5  4  1  5   16  17   0

 012  F     1 3 5 3 3 2 1 5 4 3  2  4  4  3  3  5  4  3  5  4  4  4  0  0  4  4  0  0  0  0  0  0  0   10  14   9
 012  S     1 3 4 3 1 2 1 5 3 3  2  4  1  3  4  3  5  4  3  5  4  3  4  3  5  2  3  4  0  0  0  0  0   17  11   5

 015  F     1 2 1 3 3 2 1 4 1 4  3  4  1  3  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0    7   7  19
 015  S     1 3 1 3 1 2 1 1 5 5  0  4  1  3  4  3  5  4  3  3  4  0  0  0  0  0  0  0  0  0  0  0  0   18   2  13

 018  F     1 3 5 1 1 2 1 1 5 2  3  5  1  2  3  3  5  4  2  1  4  4  2  3  5  5  3  4  2  3  1  1  3   16  17   0
 018  S     1 3 5 3 1 2 1 1 5 3  1  4  1  1  3  3  5  4  3  3  4  5  5  1  5  5  1  4  5  3  1  1  5   20  13   0

 020  F     4 3 4 5 1 4 5 5 5 3  5  5  5  2  4  3  5  2  3  4  4  5  5  5  3  4  4  4  4  5  4  5  4   11  22   0
 020  S     1 3 5 4 3 2 1 5 5 3  1  5  2  3  3  3  5  3  5  3  2  4  5  1  5  2  5  4  5  3  2  2  4   12  21   0

** KEY **   1 3 5 3 1 2 1 1 5 2  2  4  1  3  4  3  5  4  3  3  4  5  2  4  3  2  1  2  1  3  4  4  2

 021  F     1 3 5 3 1 2 1 1 5 3  1  4  1  1  4  3  5  4  3  3  4  1  2  5  3  3  3  4  1  3  1  2  3   22  11   0
 021  S     1 3 5 3 1 2 1 1 5 2  2  4  1  3  4  3  5  4  3  3  4  1  2  4  4  2  1  2  1  3  4  2  2   30   3   0

 022  F     1 3 5 5 1 5 1 4 5 3  2  4  1  3  3  3  4  4  3  2  4  3  5  3  4  5  3  4  1  5  4  3  3   16  17   0
 022  S     1 3 5 1 1 5 1 4 5 2  2  4  1  5  3  3  5  4  3  3  4  1  2  5  4  5  4  3  5  3  1  2  5   18  15   0

 024  F     1 3 5 1 3 3 1 4 2 2  2  4  1  1  4  5  5  4  3  2  4  4  2  2  5  3  1  4  3  0  0  0  0   15  14   4
 024  S     1 3 5 3 1 2 1 1 2 3  2  4  1  3  4  3  1  4  5  3  4  4  4  5  3  5  1  4  5  5  1  1  1   19  14   0
```

Figure II-8

1=CORRECT,X=WRONG,0=OMIT

FALL AND SPRING ITEM ANALYSIS CHART

RANDOM SAMPLE - GIRLS - ARITHMETIC APPLICATIONS

Item No.		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	TOTALS		
Pupil No.	F S																																		Right	Wrong	Omit
001	F	R	R	R	R	R	R	R	W	R	W	W	R	R	W	R	W	R	W	W	W	R	R	W	W	R	W	R	W	W	O	O	O	O	16	13	4
	S	R	R	R	R	R	R	R	R	R	R	R	R	R	W	R	R	R	R	W	R	R	R	R	W	W	W	W	W	R	W	W	R	W	23	10	0
002	F	W	W	W	W	R	R	R	W	O	W	O	O	O	W	W	R	R	R	R	O	O	O	O	O	O	O	O	W	W	R	O	O	O	8	10	15
	S	R	R	R	W	R	W	R	W	W	W	R	W	W	W	R	R	R	R	R	R	R	R	W	W	W	W	W	W	W	R	W	W	W	15	18	0
004	F	R	R	W	R	R	W	R	W	R	W	R	W	R	R	R	R	R	R	R	W	R	W	R	W	W	W	W	W	O	O	O	O	O	16	12	5
	S	R	R	R	R	R	W	R	R	W	R	R	R	R	R	R	R	R	R	R	R	W	R	R	W	W	W	R	R	W	R	R	W	W	24	9	0
006	F	R	R	R	W	R	R	R	W	W	O	W	W	R	R	R	W	R	W	W	W	W	W	W	W	W	R	R	W	W	W	R	W	W	13	19	1
	S	R	R	R	R	R	W	R	R	W	R	R	R	R	R	R	W	R	W	W	R	W	R	R	W	W	W	R	R	W	R	R	W	R	22	11	0
008	F	R	R	R	R	R	R	R	W	R	R	W	R	W	R	W	R	R	R	R	R	R	R	W	W	W	R	W	W	W	R	O	O	O	20	10	3
	S	R	R	R	W	R	W	R	R	W	R	R	R	R	R	W	R	W	W	O	W	R	W	R	W	W	W	R	R	W	R	R	W	R	19	13	1
011	F	R	R	R	R	R	W	R	R	W	W	W	R	R	W	W	R	R	R	R	W	W	W	W	W	W	W	W	W	W	W	W	R	W	14	19	0
	S	R	R	R	W	W	W	R	W	R	R	W	R	W	R	R	R	R	R	R	W	W	R	R	W	W	W	W	W	W	W	R	W	W	16	17	0
012	F	R	R	R	R	W	R	R	W	W	W	R	R	W	R	W	W	W	W	W	W	R	W	O	O	W	W	O	O	O	O	O	O	O	10	14	9
	S	R	R	W	R	R	R	R	W	W	W	R	R	R	R	R	R	R	R	R	W	R	W	W	W	W	R	W	W	O	O	O	O	O	17	11	5
015	F	R	W	W	R	W	R	R	W	W	W	W	R	R	R	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	7	7	19
	S	R	R	W	R	R	R	R	R	R	W	O	R	R	R	R	R	R	R	R	R	R	O	O	O	O	O	O	O	O	O	O	O	O	18	2	13
018	F	R	R	R	W	R	R	R	R	R	R	W	W	R	W	W	R	R	R	W	W	R	W	R	W	W	W	W	W	W	R	W	W	W	16	17	0
	S	R	R	R	R	R	R	R	R	R	W	W	R	R	W	W	R	R	R	R	R	R	R	W	W	W	W	R	W	W	R	W	W	W	20	13	0
020	F	W	R	W	W	R	W	W	W	R	W	W	W	W	W	R	R	R	W	R	W	R	R	W	W	R	W	W	W	W	W	R	W	W	11	22	0
	S	R	R	R	W	W	R	R	W	R	W	W	W	W	R	W	R	R	W	W	R	W	W	W	W	W	R	W	W	W	R	W	W	W	12	21	0
021	F	R	R	R	R	R	R	R	R	R	W	W	R	R	W	R	R	R	R	R	R	R	W	R	W	R	W	W	W	R	R	W	W	W	22	11	0
	S	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	W	R	R	W	R	R	R	R	R	R	W	R	30	3	0
022	F	R	R	R	W	R	W	R	W	R	W	R	R	R	R	W	R	W	R	R	W	R	W	W	W	W	W	W	W	R	W	R	W	W	16	17	0
	S	R	R	R	W	R	W	R	W	R	R	R	R	R	W	W	R	R	R	R	R	R	W	R	W	W	W	W	W	W	R	W	W	W	18	15	0
024	F	R	R	R	W	W	W	R	W	W	R	R	R	R	W	R	W	R	R	R	W	R	W	R	W	W	W	R	W	W	O	O	O	O	15	14	4
	S	R	R	R	R	R	R	R	W	W	R	R	R	R	R	R	W	R	W	R	R	R	W	W	R	W	R	W	W	W	W	W	W	W	19	14	0
026	F	W	W	W	W	W	W	R	W	R	W	W	W	W	W	W	W	W	W	W	W	W	R	W	W	W	W	W	O	O	O	O	O	O	3	24	6
	S	W	W	W	W	W	R	W	W	W	W	W	W	W	R	W	R	R	W	W	R	W	W	W	W	W	R	W	W	W	W	W	R	W	7	26	0
027	F	R	R	W	W	W	W	W	W	W	W	R	W	W	W	W	W	W	W	W	R	W	W	W	W	R	W	W	W	W	R	W	W	W	6	27	0
	S	R	R	R	W	W	R	W	R	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	R	R	W	W	W	7	26	0

Figure II-9

DATA FROM FIGURE II-8

EXPRESSED IN RIGHT, WRONG, OMIT MODE

In doing this, of course, one would have to pay attention to the fact that the data are tabulated for both fall and spring and derive a sum of "R's" for a particular item separately for the two times the test was administered.

Dividing the number of "R's" by the number of cases would give the percent correct, or the item difficulty, which has been tabulated elsewhere and commented on at some length.1/

By and large the Rights, Wrongs, Omits mode is the preferred mode of distributing the pupil responses for class use and certainly is much easier to work with in evaluating the protocols for a particular student.

Perhaps it would be helpful to conclude this discussion of the rosters of pupil responses by indicating in a summary fashion just what one would do with these data.

1. The consistency of the response from fall to spring would indeed be one of the first things to look for. (For a better evaluation of the advantages of doing this, it is suggested that the reader review the section on categorical analysis earlier in Part II.)

For instance, one can tell from the responses for pupil #001 that 14 of the 33 items were answered correctly both in the fall and spring, and therefore, assuming that none of these came about by fortunate guessing, a very large proportion of this test was nonfunctioning for this child.

However, it is most enlightening to note that there are no "RR" responses beyond item #22, so apparently at about this point in the test it seems to be a viable test for this child if measurement of change is a major goal.

There are eight instances where the choice in the fall is Wrong and the spring response is correct; i.e., the Wrong/Right combination. These are the eight items that suggest actual learning may have taken place.

2. The number Right in the spring should substantially exceed the number Right in the fall.

For example, case #008 had 20 Right in the fall but only 19 Right in the spring. Such a situation could come about if the test was highly specific to the curriculum of the grade below and the child had not at all been exposed to the content of the current curriculum at the time he took the test in the fall.

This would heighten his opportunity to improve his score as the result of seven months of instruction; but apparently the original score (or the final score) was not valid, since there is an actual loss!

Case #002 would appear to be a case falling in this category, but an examination of the individual responses item by item makes one wonder. In the last 13 items in the test there are only 3 Right in the spring, and all of the remaining responses (that is, 10) are Wrong; so it is patently evident that guessing has occurred in this particular instance.

3. There are other ways of studying these data, limited only by a person's imagination and actual knowledge of the case.

A teacher examining data of this sort, knowing the child and knowing his day by day performance, can find this kind of exercise enormously illuminating. This can result in a decision to consider a test invalid, so far as the total score is concerned (if a pattern like case #002 is found), largely due to the very erratic types of responses to be found in the spring compared to fall.

4. To generalize broadly, in a non-guessing situation the Right responses will constitute a very large majority of the items attempted, with few Wrongs and Omits in indirect proportion to attempts.

Where guessing is rampant, a pupil's total score will approach the average random score - with considerable chance variation in both directions.

Where some knowledge is present and some item response is marked most of the time (i.e., very few Omits or none at all), one must proceed cautiously. Data from the predicted score analysis and from the R/A analysis will help, but there is NO infallible way of identifying which "R" responses are guesses and which are the result of learning.

1/ See pages 8 to 14, Part II.

SAMPLE CASES FOR ILLUSTRATIVE PURPOSES

We have selected some sample cases drawn from the actual roster of pupils for the item analysis made in connection with this study. On the Item Analysis Data sheet the recording of the choice made by the pupil (i.e., 1-2-3-4-5 or a-b-c-d-e, whichever it might be) has been changed to the Right-Wrong-Omit mode of recording item data without regard to the alternative chosen.

Thus it is possible, without the use of a key, to count the number Right for any segment of the child's item analysis response pattern. Rights divided by Rights plus Wrongs gives the guessing ratio (as it was originally called), which of course is the ratio of Rights over Attempts (R/A).

We have chosen three samples from the Random Sample population. In addition to the specific item data noted above, all other available information concerning each of these three cases has been collected and considered so as to give as complete a picture of each pupil as possible. These data have been recorded on the Individual Profile Chart and Personal Data Sheet for each case.

The answer sheets for these children for both spring and fall are available and they have been examined for any departure from an acceptable method of marking.

We have taken a quick look at each child's performance on all of the tests he took as it is "on view" on his answer sheet. We found nothing that looked atypical; that is, nothing that would say "Stop" to the computer under regular scoring routines for any of the cases.

Each child's school learning potential, or IQ as derived from the Otis-Lennon Mental Ability Test: Elementary II: Form J, has been checked and is considered along with other data. The sex of each child in this analysis has been noted, although this appears to be of little significance, according to our analyses of the data as a whole.

We know the child's birthdate and his testing date (and, therefore, his chronological age), and with this information plus the score on the Otis-Lennon Test we have computed the Deviation IQ by looking it up in the appropriate tables provided by the publisher to check the information already written down for each pupil.

Unfortunately, we do not have the advantage of seven months of almost daily observation of each child; i.e., every day of the week except Saturdays and Sundays for 140-150 days. This is the big advantage of the classroom teacher's observation, and it would be foolhardy indeed for anyone interpreting test scores ever to ignore it. Even the observation for the short period of time prior to the administration of the tests in mid-October is very valuable, especially if an appropriate system of cumulative records is in force.

Information concerning the children moving up from the lower grades should always be passed along to the next teacher at the beginning of the school year - objective information, especially, as well as observational evaluations.

At the upper grades this more often is done formally through a cumulative record card, but such information is infrequently passed on from level to level; i.e., elementary to junior high to senior high, etc. Computer technology has done much to change this omission in places where it is available, like Dade or Pinellas Counties in Florida and hundreds of other large city and county units.

It will be recalled that part of our analysis has been done by stratifying the data, not only in the usual ways but also by separating the sample into a group of those who do not attempt all of the items (i.e., follow the supposedly classic pattern - see Sample A) on the one hand and, on the other hand, into those who attempt all of the items, regardless of how many they answer correctly (see Sample B); i.e., we show the two extremes of a continuum and not a dichotomy.

A peripheral value of having the item analysis presented in the Rights-Wrongs-Omits mode is that it not only allows the teacher to get the general pattern of a single pupil's responses item-by-item for comparison with the items as presented in the test booklet, but by summing the columns for the class rosters presented in this fashion one also attains the number of Rights, Wrongs, and Omits for the class as a whole.

This information both for the individual and for the class, especially with respect to those items that have been answered incorrectly or omitted (or have been answered inconsistently from fall to spring, in this particular study), provides information that is surely as worthwhile as the total score on the tests interpreted in terms of any norm, whether local or national.

We show this type of analysis for all three sample cases, but it certainly is too laborious for the teacher to do for all children. It is entirely feasible if computer assistance is available. We have previously shown a sample page for one test in class roster form (see Figure II-9).

In any contained system (i.e., a response system in which a child chooses his answer from a selected number of alternatives) there is the possibility of a chance response, so a Right response either in the fall or in the spring, or even in both fall and spring, is not incontrovertible evidence that the child knows the knowledge or skill measured by the particular item.

It is perfectly practical and desirable to assume that a Right response in the fall is more likely to indicate knowledge (i.e., evidence of "knowing" rather than a chance response, especially in the early, easier items), but a consistent Right response in the spring leads more convincingly to the conclusion that the child does, indeed, know the skill or information measured by a particular item. It is too late to wait for that if the tests are to be used during the "between-testing" period for improving instruction.

If his reply is inconsistent (i.e., fall "R", "W" spring) his first "R" response was probably a guess.

However, the only way that a teacher can ever know whether a particular knowledge or skill is really mastered is to observe the application of the child's knowledge or skill in a whole series of everyday situations where that knowledge or skill is essential for success on a particular task.

To know whether the child truly knows the number combination of 8+9 "for sure," he must be put to the test in a number of varied situations where a basic segment of the total task requires that he know and apply the knowledge that 8+9=17.

This kind of information is beyond our knowing on the basis of the test data available, even in a fall-spring testing program.

Here again, the importance of the teacher's constant reevaluation of the knowledges and skills of her pupils in a repetitive and "maintenance of skills" fashion is perfectly evident. This kind of approach, when done by an appropriate test, would be what is more widely known now as "criterion reference testing"; i.e., no norm is supposedly required, although this is more a "seemingly so" situation than an actual one.

Let it be taken for granted, in this particular instance, that the primary purpose of the testing program such as the one presently considered is to improve instruction. The testing program does this by providing objective evidence that material presented either has been known at one time or not known. (R versus W)

If not known, it has to be learned during the course of instruction between first testing and second testing if the response is "WR" or "OR" - the only categories that really measure.

It assumes that the teacher will take the evidence of the first test, when summed overall for the class as a whole, to indicate areas of weakness which need to be strengthened. When considering the pattern of responses for all items on a particular test given in the fall for a particular child, the teacher will try to assess his status, identify areas of weakness, and modify and strengthen his instruction at certain points so as to build on what the child does know and to provide the support and help he needs to learn what he should know in accordance with the local curriculum.

With this background, let us consider now the cases that have been selected in the manner indicated above case by case. Each case data are presented on a separate page which contains all of the available information about a particular child, but these pages are run into the text in such a way that a discussion of a particular child immediately follows the presentation of all the available data concerning that child.

SAMPLE A

INDIVIDUAL PROFILE CHART AND PERSONAL DATA SHEET

New Hampshire Statewide Testing Program
Otis-Lennon Mental Ability Test: Elementary II: Form J - Fall 1969
Stanford Achievement Test: Intermediate I: Form X - Fall 1969 and Spring 1970

GRADE 4

Case # 452 RANDOM SAMPLE ✓ TITLE I___ Boy ✓ Girl___

School: Public___ Parochial ✓ City or Town NASHUA

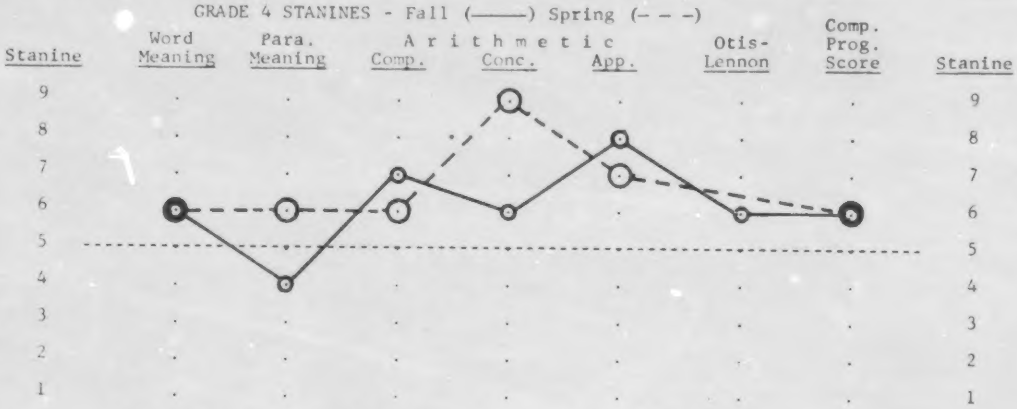
Date of Fall Testing 10/13/69 Date of Birth 4/4/60 Age: 9 years 6 months

Median Grade 4 Age, Fall 1969: 9 years 6 months - Random Sample ✓ Title I___

Norms: OLMAT - National - DIQ 107 Percentile Rank: Age .67 Grade .71 Stanine: Age 6 Grade 6

SAT - State - Random Sample ✓ Title I___

GRADE 4 STANINES - Fall (———) Spring (– – –)



	F	S	F	S	F	S	F	S	F	S	Fall	F	S
Stanines	6	6	4	6	7	6	6	9	8	7	6	6	6
RIGHTS	19	25	19	36	17	22	17	27	20	22	42		
Wrongs	5	4	10	5	2	6	6	5	4	6	8		
Omits	14	9	31	19	20	11	9	0	9	5	30		
No. of Items	38		60		39		32		33		80		
Rights/Attempts (R/A)	.79	.86	.66	.88	.89	.79	.74	.84	.83	.79	.84		
Pred. Score	19.5	29.2	28.4	41.6	19.9	21.7	22.5	28.8	22.4	23.6			

SAMPLE A

Item Analysis Data

RANDOM SAMPLE - Boys - Case #452

```
                                  1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 3 3 3 3 3 3 3 3 3 3 4
Item Number -   1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0    R   W   O

OLMAT: EL. II: J   DIQ=107   Fall
Otis-Lennon
(80 5-choice)
 Items 1 - 40:  R R R R R|R R R R R|R R Ẇ R R|R R Ẇ R R|R R R R R|Ẇ R Ẇ R R|R R R R R|R R Ẇ R Ẇ|

 Items 41 - 80: R R R R R|R Ẇ Ẇ R R|O O O O O|O O O O O|O O O O O|O O O O O|O O O O O|O O O O O|  42   8  30

SAT: INT. I: X
Word Meaning
(38 4-choice)
   Fall   -     R R R Ẇ|R R R R|R O Ẇ R|R R R Ẇ|R Ẇ R R|R R Ẇ R|R O O O|O O O O|O O O O|O O       19   5  14
   Spring -     R R R R|R R R R|R R R R|R R R R|R R R R|Ẉ R R R|R R Ẉ Ẉ|Ẉ O O O|O O O O|O O       25   4   9

Paragraph Meaning
(60 4-choice)
 Items 1 - 40:
   Fall   -     R R R R|Ẇ Ẇ Ẇ R|R R R Ẇ|R R Ẇ R|R R R R|R R Ẇ Ẇ|Ẇ Ẇ R R|Ẇ O O O|O O O O|O O O O|
   Spring -     R R R R|R R R R|R R R R|R R R R|Ẉ R R R|R R R Ẉ|Ẉ Ẉ R R|R R R R|Ẉ R R R|R R R R|
 Items 41 - 80:
   Fall   -     O O O O|O O O O|O O O O|O O O O|O O O O|                                          19  10  31
   Spring -     R O O O|O O O O|O O O O|O O O O|O O O O|                                          36   5  19

Arithmetic Computation
(39 5-choice)
   Fall   -     R R R R R|R R R R R|R O Ẇ R R|R R R R Ẇ|O O O O O|O O O O O|O O O O O|O O O O     17   2  20
   Spring -     R R R R R|R R R R R|R Ẉ R R Ẉ|R Ẉ Ẉ R R|R O R R O|Ẉ O R R O|Ẉ R O O O|O O O O     22   6  11

Arithmetic Concepts
(32 4-choice)
   Fall   -     R R R R|R Ẇ R R|R R R Ẇ|R R Ẇ Ẇ|R R R R|Ẇ R Ẇ O|O O O O|O O O O|                  17   6   9
   Spring -     R R R R|R R R R|R R R R|R R R R|R R R R|R R R Ẉ|R R Ẉ R|R Ẉ Ẉ Ẉ|                  27   5   0

Arithmetic Applications
(33 5-choice)
   Fall   -     R R R R R|R R R R Ẇ|R R Ẇ R R|R R R R R|R Ẇ R Ẇ O|O O O O O|O O O                 20   4   9
   Spring -     R R R Ẉ R|R R R R Ẉ|R R R R R|R R R R R|R Ẉ R R R|Ẉ Ẉ Ẉ O O|O O O                 22   6   5
```

SAMPLE A

Our first case is a boy, deliberately chosen from the random sample for reasons that will appear shortly. His age was 9 years and 6 months as of the date of testing in the fall; exactly the median age of the group.

He had an Otis-Lennon DIQ at that time of 107. His Otis-Lennon percentile rank *according to age* was .67, which corresponds to a stanine of 6. His *grade placement* percentile rank was .71 on the same test, and it also corresponds to a stanine of 6

Thus we have a youngster who is exactly at age for grade but who is a little brighter than the average *in terms of measured mental ability*; i.e., a little better performance should be expected of him, all things considered.

Interpretation

His school performance, as shown on the Individual Profile Chart for the five tests in which we are presently interested, indicates that he earns stanines which generally are in the 6 or 7 range in the fall with one 8, namely in Arithmetic Applications.

He also has a *4* stanine in the fall in Paragraph Meaning. He is probably significantly below the reading grade level of the fall random sample of children tested in 1969.

This is a rather unusual situation in light of his Otis-Lennon stanine of 6. There is reason to suspect that perhaps it may be a true reflection of his situation in view of a very substantial gain in Paragraph Meaning during the seven months between tests, hopefully due to some successful remedial instruction.

The Personal Data Sheet also gives fall and spring raw scores (Rights), the number of Wrongs, the number of Omits, the number of items in each test, and the R/A ratio.

This case is most notable for the number of Omits, reflected in the relatively very high R/A ratios. This use of the omit technique, rather than random guessing, is convincing preliminary evidence that the test data are valid.

The earned stanines are shown on the plotted profile. The stanines are based upon the distributions of scores for the *random sample* and were computed separately for fall and spring. The stanines used in the fall and based on the fall random sample are shown as a solid line on the profile; the spring stanines are shown as a broken line. Thus growth or change is reflected in deviations from the stanine average (5) from subject to subject.

For example, a consistent upward trend in such an instance indicates more than average growth relative to the median of the conversion sample; a downward trend, the opposite.

One other item of information which looks interesting (and to some extent suggests a problem area) is in the Arithmetic Applications data, where he makes a score of 20 in the fall but has a gain of only 2 points, to 22, in the spring.

However, the score of 20 gives him a stanine of 8 in the fall random sample. He probably already had been exposed to and had learned, to a very substantial degree, most of the material that was presented through the fourth grade. (In or out of school? A transfer student? Naturally gifted in the number area?)

His failure to make a gain in score of appreciable amount in Applications, though he gained substantially in Concepts, might very well be due to the fact that he did get very much additional exposure to problem-type material consistent with his ability to perform as indicated by the stanine of 8 in the fall.

This high math score and corresponding stanine versus low reading score and corresponding stanine in the fall also is a common indication of a reading difficulty.

He has quite apparently used the omit technique generously as a "don't know" indicator in every test both fall and spring with the exception of the spring Arithmetic Concepts Test, where he omitted *no* items but still came up with an R/A ratio of .84.

This is not too surprising in view of the fact that there are only 32 items in this test anyway, and his original score was 17 Right and 6 Wrong. In the spring, he had only 5 Wrong and 27 Right, for a gain of an extraordinary 10 points.

Comparing his predicted and actual scores, we see that they are not only high but generally fairly close (again with the exception of Paragraph Meaning), all of which supports the conclusion that his reading was a problem area in the fall.

This possibility of a correctable reading deficiency is great in view of the fact that he makes an enormous raw score gain in Paragraph Meaning, from a raw score of 19 to a raw score of 36 in the spring (stanine gain, relative to separate sets of fall and spring stanines, of 2) while he improved the R/A ratio - .66 in the fall to .88 in the spring. It is the latter piece of information that is most convincing.

On the whole, this child is not one who is going to give anybody any trouble in school in terms of his subject matter oriented performance.

His profile in the spring is remarkably uniform in comparison with his measured mental ability, with the noted exception of Concepts where he has exceeded the expected stanine substantially. Everything else is within chance limits of his Otis-Lennon grade stanine of 6.

The effect of regression, it must be noted, in above average ability is to maintain the status quo or regress toward the mean; he did not regress.

The matter of making a test profile for any child is one of great concern. It shows a great deal graphically if the profile is in comparable units. Hence the scores have been profiled in separate stanines for fall and spring, which are comparable because the group is comparable; i.e., identical, in fact. This point is rather subtle but of great significance.

Saying this makes it necessary for us to try to clarify the idea back of this method of indicating change or inconsistency in growth pattern.

This writer has long advocated profiles in such comparable stanines as a way of reflecting growth rather than measuring it directly by the magnitude of a change in some kind of standard score.

A direct measure of growth has long been sought as a highly desirable statistic, but this has proved to be almost impossible to achieve in any kind of continuous standard scores because the continuous growth curve (i.e., the line of relation drawn through medians or means) varies in slope from subject to subject.

Any set of scaled scores that attempts to do what Thurstone's scaled scores are supposed to do, namely create a kind of artifical absolute zero and to scale the scores along a continuum from the very beginning grades to the highest possible grade, is doomed to failure as a measure of comparable growth unless the growth curve is the same in all areas. Furthermore, the growth potential of a child is just not going to be the same from subject to subject or from one grade to another, for reasons too numerous to mention.

In a subject such as Word Meaning, growth is very subject to influences from the total outside environment as well as from in-school instruction; while in another subject (such as Arithmetic Applications or Problem Solving) is very largely a school-oriented skill with, generally, little or no outside incidental learning.

Fundamentally, the idea here is that a child's "growth" (i.e., tested development) is reflected by the extent to which he deviates from the average of his peers and/or from his own average from year to year allowing for random or chance errors (standard error of measurement).

In addition to the tests which we are including in our profile, we have added one more statistic; namely, a Composite Prognostic Score based upon weighted stanines.

Such a composite is by far the most stable value of any other single stanine score. Obviously, because it will be made up by a weighted average of the stanines within the total number of tests of achievement plus the measured mental ability test, the item base is much greater. This makes for a more reliable individual pupil reference point.

Weights can be assigned to the tested elements either by statistical methods or by judgment. In this case, the weights used were judged weights very similar to those used for years in the writer's New Hampshire statewide 8th grade programs and other similar programs.

Case #452 has stanine composites of 6 for both fall and spring, using weights as listed below: 1/

OLMAT Raw Score	3
SAT Paragraph Meaning	2
" Arithmetic Computation	2
" Word Meaning	1
" Arithmetic Concepts	1
" Arithmetic Applications	1
	10

1/ The Composite Prognostic stanines for this individual were obtained by averaging his separate stanines; in general practice the sum of the weighted stanines are re-scaled to avoid the shrinking effect of an averaging procedure.

Constant Failure as a Personality Determinent

Another child, perhaps a slow-learning child with stanines running in the 2-3-4 range, all too often will experience the constant stigma of failure because he will be at the low end of the stanine scale (for shame!) and, as a result, will rapidly develop a negative attitude toward school and toward his own learning potential.

Test results consistent with potential can never be considered to be evidence of failure. Therefore by tracking a child from year to year, using his own weighted stanine position based on both mental ability plus measures for achievement to obtain a Composite Prognostic Score, we are able to see to what extent he varies from year to year on an empirical basis and thus grasp more firmly the type of individual we are dealing with.

Of all the scores reported, the Composite Prognostic Score gives the most practical single estimate of what could be expected from this pupil barring some traumatic changes in some aspect of his situation. The results in this case bear out this contention.

Let us abandon all talk of success or failure where test scores are involved and we will be well on our way toward obtaining acceptance from the child of what he is as regards verbal learning and without stigma, because it is no one's opinion but a reflection of facts!

This assumes the development and acceptance of the practices and attitudes advocated in this report, including especially the reality of great individual differences.

SAMPLE B

INDIVIDUAL PROFILE CHART AND PERSONAL DATA SHEET

New Hampshire Statewide Testing Program
Otis-Lennon Mental Ability Test: Elementary II: Form J - Fall 1969
Stanford Achievement Test: Intermediate I: Form X - Fall 1969 and Spring 1970

GRADE 4

Case # 101 RANDOM SAMPLE ✓ TITLE I__ Boy ✓ Girl__

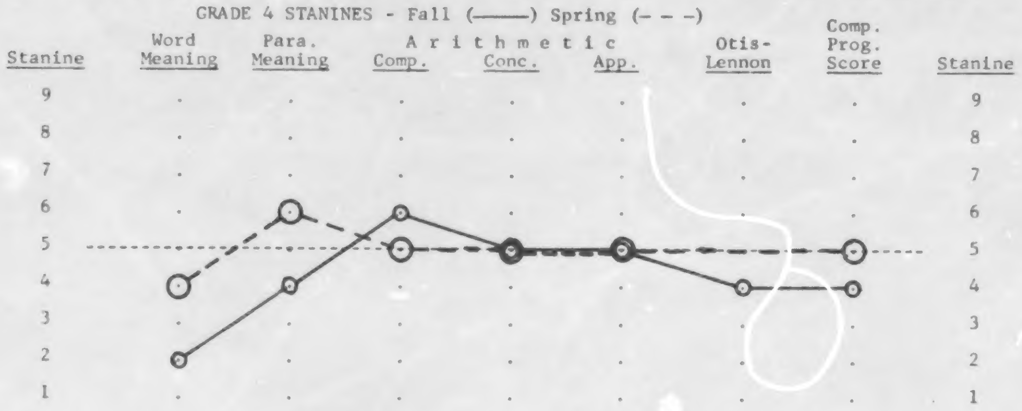
School: Public ✓ Parochial__ City or Town DOVER

Date of Fall Testing 10/20/69 Date of Birth 12/1/60 Age: 8 years 11 months

Median Grade 4 Age, Fall 1969: 9 years 6 months - Random Sample ✓ Title I__

Norms: OLMAT - National - DIQ 100 Percentile Rank: Age .50 Grade .36 Stanine: Age 5 Grade 4

SAT - State - Random Sample ✓ Title I__



	F	S	F	S	F	S	F	S	F	S	Fall	F	S
Stanines	2	4	4	6	6	5	5	5	5	5	4	4	5
RIGHTS	6	17	19	36	14	18	12	15	14	16	26		
Wrongs	32	21	41	24	25	21	20	17	19	17	54		
Omits	0	0	0	0	0	0	0	0	0	0	0		
No. of Items	38		60		39		32		33		80		
Rights/Attempts (R/A)	16	.45	.32	.60	.36	.46	.38	.47	42	.48	.33		
Pred. Score	5.2	11.7	16.7	36.1	12.9	16.3	12.6	14.4	12.4	13.1			

SAMPLE B

Item Analysis Data

RANDOM SAMPLE - Boys - Case #101

```
                                   1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 3 3 3 3 3 3 3 3 3 3 4
Item Number - 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0     R   W   O

OLMAT: EL. II: J   DIQ = 100   Fall
Otis-Lennon
(80 5-choice)
Items 1 - 40  R R R R R|R R R W R|R W W R W|W R W W W|W W W W W|W W W W W|R W R W R|W R W R W|

Items 41-80   W R W W R|W R W W W|R R W W W|W W W W R|W W W W R|W W W R W|W W W W R|W W W W W|  26  54   0

SAT: INT. I: X
Word Meaning
(38 4-choice)
   Fall   -   R R W W|W W W W|W W W R|W W W W|W R W W|R W W W|W W W W|W W W W|W R W W|W W           6  32   0
   Spring -   R R R W|R W W W|R W W W|R W W W|R R W W|R R R W|R W R R|R R W W|W R W W|W W          17  21   0

Paragraph Meaning
(60 4-choice)
 Items 1 - 40:
   Fall   -   W R W W|W W W R|W R R W|W W W R|W R R R|W R R W|R W R W|W W W R|R W W R|W W W R|
   Spring -   W R R R|R R R R|R R R R|R R R R|W R R W|R R W R|W W R W|W R W R|R R W R|R R R W|
 Items 41- 60:
   Fall   -   W R W W|W R W W|W W W W|W W W W|R W W W|                                             19  41   0
   Spring -   W R R W|W W W R|W W R W|W R W R|W W W R|                                             36  24   0

Arithmetic Computation
(39 5-choice)
   Fall   -   R R W R R|R R R W W|W W R W R|R R W W W|W W W W R|R W W W W|R W W W W|W W W W      14  25   0
   Spring -   R R R R R|W R W R R|R W R W R|W R W W W|W W R W W|W W W W R|R R W W W|W W R R      18  21   0

Arithmetic Concepts
(32 4-choice)
   Fall   -   R R R R|W W W W|R W W W|W R R W|R R W W|R W W R|W R W W|W W W W|                    12  20   0
   Spring -   R W W R|W W W R|W R R R|W R R W|R W R W|W R W R|W R W R|W W R W|                    15  17   0

Arithmetic Applications
(33 5-choice)
   Fall   -   W R R R W|W R R R R|W R W W R|W W R W W|R R W W R|W W W W R|W W W                   14  19   0
   Spring -   R R R W R|W R R R W|W R R W R|W W W W W|R R R R R|W R W W W|W W W                   16  17   0
```

SAMPLE B

Our second sample case is summarized on the Individual Profile Chart and Personal Data Sheet preceding this interpretation.

This case was drawn from the random sample and is a boy going to public school who, at the time of fall testing, was 8 years and 11 months old compared to the median age for the grade of 9 years and 6 months. His Otis Deviation IQ was 100. His percentile rank on the Otis was .50 on an age basis and on a grade basis only .36.

Note first the discrepancy in the chronological age. This child is about 7 months younger than the average age in the grade, but his intelligence level as recorded is about typical for the community in question.

Seven months minus difference at the fourth grade level in terms of chronological age, and in this case quite unquestionably an equal or greater difference in mental age, can make a substantial difference in achievement.

This child must have been admitted to school about as early as the law would allow. Without question, he would be better off if he were in grade 3 rather than in grade 4; all subsequent data support this conclusion.

At this point, with all the evidence in hand, we must ask why he was ever allowed to get into 4th grade! A true ungraded primary system would have certainly found it highly desirable to give him at least four years to complete grade 3, and possibly even five years!

Turning now to the data at the bottom of the profile sheet, we see first one very notable element; namely, that he has omitted no items on any test including the Otis-Lennon.

In other words, he has immediately indicted himself as a guessing person since, with this chronological age and this level of mental ability, he could not possibly be working effectively in the latter part of any Stanford Test, all of which have been shown to be difficult for the average child in the state as a whole, and certainly much too difficult for the younger children in grade 4.

Looking first at his earned score on the Otis-Lennon Test, we see that he received a score of 26 Right out of the 80 questions, which is only 10 points of score above the chance level. This immediately raises the question as to how he could get an IQ of 100 with such a relatively low score.

The answer must lie in the fact that he was taking a level of the Otis-Lennon that was too hard for him by virtue of the fact that he was in the fourth grade in spite of his being underage for the grade, and the Otis-Lennon level used was one which was recommended for use at the fourth grade level but this was the lowest grade at which it should be used.

Furthermore, the directions do not make any allowance whatsoever for the influence of guessing. A glance at the pattern of his responses on the Otis-Lennon, as shown on the Item Analysis Data sheet, indicates that he answered a few items at the beginning of the test correctly and then began a Right-Wrong type of response which degenerates at about #17 into a pattern which could be accounted for almost wholly by random marking without reference to the test booklet at all.

It is possible that he really only answered about twelve questions on the basis of knowledge, and the remainder of the Rights are largely due to chance. A score of 12 in conjunction with his age would yield an IQ of only 77, and not 100. This is probably an underestimate of his mental ability level, but it certainly is strong evidence that the 100 is too high.

For example, he does get an occasional item correct well along toward the end of the test, the most outstanding example being item #75. However, this is preceded by a string of five Wrong responses and the five subsequent items are all answered incorrectly as well.

The Otis-Lennon is an 80-item test with five alternatives and no specific warning against guessing. The average chance score, therefore, on the test is 1/5 of the total number of items, or 16, and his earned score of 26 falls only 10 points above this average chance level.

It would be entirely within the realm of possibility for him to have gotten a score of as high as 26 without ever looking at the test booklet whatsoever, but simply marking the answer sheet; but the fact that he did answer a sequence of items at the beginning of the test correctly is convincing evidence that this certainly was not the case.

His percentile ranks on the Otis, both on the basis of the age group to which he belongs and the grade group, are based upon the score of 26 on the assumption that this is a valid score. Even so, he achieves a percentile rank of only .50 when compared with other children of the same age, who would not typically be in fourth grade and certainly not if their performance during the first three grades was what could very well be anticipated it was from the data we have at hand.

His percentile rank according to grade is only .36, meaning that his score of 26 is reached or exceeded by 64% of children in the fourth grade in the national standardization population on which the Otis norms are based.

A Consideration of Rights, Wrongs, and Omits

Looking at the data provided at the bottom of the profile page, we see first of all that the Right scores are low, not only on the Otis but for all of the tests in the Stanford Battery. In fact, the only instance where the number of Rights exceeds the number of items answered incorrectly is in the spring of the year, when he answered 36 items Right in Paragraph Meaning and answered incorrectly 24.

His gains from fall to spring are reasonably good. In fact, the gain from 19 to 36 in Paragraph Meaning, if it could be taken literally, would be an astonishingly high gain, and his gain in Word Meaning from 6 to 17 is hardly less surprising.

Rights/Attempts

As indicated in the text, the R/A ratio shows the proportion of all items attempted which were answered correctly. As one would expect from the data previously presented, he tends to be under the median values for fourth grade children, and mostly by substantial amounts.

Actually, he does not exceed the median in any instance either fall or spring, but his ratios tend to be better in the math field in the spring than they were in the fall. This is also a group tendency, strengthened by the fact that these pupils had been studying related material for a period of seven months, and therefore by some amount had reduced the opportunities to guess by their actual increment in knowledge.

In the absence of any other information, one would conclude that this child had made rather substantial progress in both vocabulary and reading during the seven months between testing and that his gains in the Arithmetic area, although small, have to be interpreted in view of the fact that the gains for the state as a whole also were small.

At this point, we are led to raise the hypothetical question: Where did this child pick up the "Attempt-All" pattern of response? Was it early, in the attempt to live up to a role in which he was quite unwittingly cast by being admitted at such an early age?

One must further wonder to what extent his performance in class was comparable to his performance on the tests that he took in the Stanford plus the Otis.

In other words, according to the teacher's observation did he appear to read fairly well? Was his seatwork in arithmetic reasonably good? Or, on the other hand, did the teacher perceive him as being essentially a slow-learning child? Was there any recognition of the fact of his being underage as well as probably below average in mental ability, if one allows for guessing?

Summary

After a careful examination of the test information, taking into account the proclivity of this child to mark all responses regardless of knowledge and his generally poor R/A ratios, we have to conclude that his tests were substantially invalid as measures of his true status both in the fall and in the spring, although the tests do suggest some rather amazing improvement in the language area during the course of the year.

The true nature of this child's performance is really seen best in the summary of his item by item responses, as we see the pattern of chance responses emerging clearly after a very few of the easiest items have been answered.

The Predicted Score

As for all cases analyzed, we used the formula Predicted Score = A + (A'/C x D) to predict this pupil's scores on the Stanford Achievement Test. 1/ These appear in the last line of the Personal Data Sheet.

For every test except Word Meaning in the spring, which in itself is a curious situation, the proportion of all items marked Right seemed adequate to make this prediction reliable. However, from the previously established fact that guessing is a "way of life" for this child, we know that even here his Rights scores for the first 20 (23) items probably are inflated in most

1/ See pages II-34 and 35 for further explanation of this procedure.

tests. (See Item Analysis Data following the Individual Profile Chart and Personal Data Sheet.)

In Word Meaning, only four responses in the first twenty are correct, and one of these is followed by a Wrong response in the spring.

In Paragraph Meaning, there are ten Rights in the first 23 in the fall, but two of these have a "W" response in the spring.

To take one more instance, in Arithmetic Concepts nine out of the first twenty fall responses are Right, but of these four are followed by Wrongs in the spring!

In spite of all this, the agreement of total Rights (by machine) and our predictions are not off badly. Of the ten predictions (fall and spring), eight are lower than the actual score, one spring prediction is the same as his earned score, and in one his fall prediction is fractionally higher.

This is the type of pattern expected of a guessing child; i.e., prediction lower than machine scores. The first twenty items are the _easy_ items, where guessing is _less_ necessary because it is actually easier (or more satisfying psychologically) to answer out of knowledge than to guess; while later items of increasing difficulty are impossible to answer, except by guessing, in almost every instance.

One must conclude on every basis that this child's performance on this test should be completely disregarded as a valid measure of his knowledge, generally overestimating by substantial margins what he is truly capable of doing and making it very desirable to throw out the results totally.

Perhaps the most significant thing that can be said about this child is the fact that if one were to deal solely with total scores or with the stanine profile, entirely erroneous conclusions could be drawn.

It is only when one notes that there are _no_ omits, and then actually looks at the list of item responses, that the conviction that his test result is invalid grows so strong as to make it necessary to declare the case totally erroneous and actually a detriment to the child to be retained in his record.

Answer Sheet Study - II

SAMPLE C

INDIVIDUAL PROFILE CHART AND PERSONAL DATA SHEET

New Hampshire Statewide Testing Program
Otis-Lennon Mental Ability Test: Elementary II: Form J - Fall 1969
Stanford Achievement Test: Intermediate I: Form X - Fall 1969 and Spring 1970

GRADE 4

Case # 033

RANDOM SAMPLE ✓ TITLE I__

School: Public ✓ Parochial__ City or Town CLAREMONT Boy__ Girl ✓

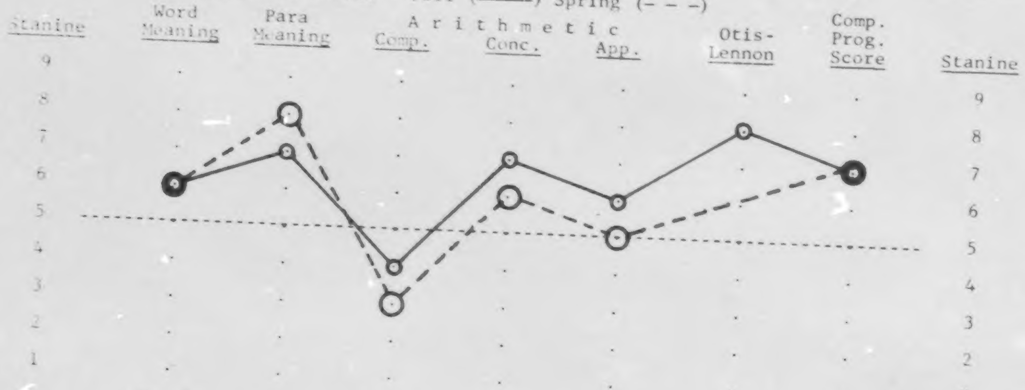
Date of Fall Testing 10/13/69 Date of Birth 9/18/60 Age: 9 years 1 months

Median Grade 4 Age, Fall 1969: 9 years 6 months - Random Sample ✓ Title I__

Norms: OLMAT - National - DIQ 128 Percentile Rank: Age 96 Grade 93 Stanine: Age 9 Grade 8

SAT - State - Random Sample ✓ Title I__

GRADE 4 STANINES - Fall (———) Spring (- - -)



	Word Meaning F	Word Meaning S	Para Meaning F	Para Meaning S	Arith. Comp. F	Arith. Comp. S	Arith. Conc. F	Arith. Conc. S	Arith. App. F	Arith. App. S	Otis-Lennon Fall	Comp. Prog. Score F	Comp. Prog. Score S
Stanines	6	6	7	8	4	3	7	6	6	5	8	7	7
RIGHTS	19	25	36	46	9	10	19	21	17	15	57		
Wrongs	8	6	10	14	19	19	6	11	9	18	15		
Omits	11	7	14	0	11	10	7	0	7	0	8		
No. of Items	38		60		39		32		33		80		
Rights/Attempts (R/A)	.70	.81	.78	.77	.32	.34	.76	.66	.65	.45	.79		
Pred. Score	22.1	24.7	35.1	41.6	8.2	9.5	21.0	20.1	18.6	14.4			

SAMPLE C

Item Analysis Data

RANDOM SAMPLE - Girls - Case #033

```
                                1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 3 3 3 3 3 3 3 3 3 3 4
Item Number - 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0   R   W   O

OLMAT: EL. II: J   DIQ = 100   Fall
Otis-Lennon
(80 5-choice)
Items 1 - 40: R R R R R|R R R Ẇ R|R R R R R|R R R R Ẇ|Ẇ R R R R|R R Ẇ R R|R Ẇ R R R|R R R R Ẇ|

Items 41 - 80:R R Ẇ R R|Ẇ R Ẇ R R|R R Ẇ R R|R R R R R|Ẇ R R Ẇ Ẇ|R R Ẇ R R|Ẇ R O O O|O O O O O| 57  15   8

SAT: INT. I: X
Word Meaning
(38 4-choice)
   Fall   -   R R R R|R Ẇ R R|Ẇ R R R|R R R R|Ẇ R R R|Ẇ Ẇ Ẇ R|Ẇ R Ẇ O|O O O O|O O O O|O O      19   8  11
   Spring -   R R R R|R R R R|R R R R|R Ẉ R R|R R R R|R Ẉ Ẉ R|Ẉ R R R|R Ẉ Ẉ O|O O O O|O O      25   6   7

Paragraph Meaning
(60 4-choice)
 Items 1 - 40:
   Fall   -   R R R R|R R R R|R R R R|R R R Ẇ|R Ẇ R R|R R R Ẇ|Ẇ Ẇ R R|R R R R|Ẇ R Ẇ Ẇ|R R R R|
   Spring -   R R R R|R R R R|R R R R|R R R R|R R R R|R R Ẉ Ẉ|Ẉ Ẉ Ẉ Ẉ|R R R R|Ẉ R R R|R R R R|
 Items 41 - 60:
   Fall   -   R R R Ẇ|R Ẇ O O|O O O O|O O O O|C O O O|                                         36  10  14
   Spring -   R R R R|R Ẉ R Ẉ|R Ẉ R R|R Ẉ R R|Ẉ Ẉ Ẉ R|                                         46  14   0

Arithmetic Computation
(39 5-choice)
   Fall   -   R R Ẇ R R|Ẇ Ẇ Ẇ Ẇ Ẇ|O Ẇ R Ẇ R|Ẇ R Ẇ Ẇ Ẇ|Ẇ O Ẇ Ẇ O|Ẇ O R O O|Ẇ O O O Ẇ|O Ẇ O R     9  19  11
   Spring -   Ẉ R R Ẉ R|R Ẉ Ẉ Ẉ Ẉ|R Ẉ Ẉ Ẉ Ẉ|Ẉ R R Ẉ R|R O Ẉ Ẉ O|R O O Ẉ Ẉ|Ẉ Ẉ Ẉ O O|O O O O    10  19  10

Arithmetic Concepts
(32 4-choice)
   Fall   -   R R R R|R Ẇ R Ẇ|R R R R|Ẇ R R Ẇ|R R R Ẇ|R R R Ẇ|R O O O|O O O O|                 19   6   7
   Spring -   R R R R|R Ẉ R Ẉ|R Ẉ R Ẉ|Ẉ R R R|R R R Ẉ|R Ẉ R Ẉ|R R Ẉ R|R Ẉ R Ẉ|                 21  11   0

Arithmetic Applications
(33 5-choice)
   Fall   -   R R R R Ẇ|R R Ẇ R R|R R R R R|Ẇ Ẇ R R Ẇ|Ẇ Ẇ Ẇ Ẇ R|R O O O O|O O O                17   9   7
   Spring -   R R R R Ẉ|Ẉ R Ẉ Ẉ Ẉ|Ẉ R Ẉ R Ẉ|R R R R Ẉ|Ẉ R Ẉ R Ẉ|Ẉ Ẉ Ẉ Ẉ R|R Ẉ Ẉ                15  18   0
```

SAMPLE C

This third and last case we will discuss is very interesting because it is so different from the two previous cases or from a typical profile pattern.

It is a girl in a small city. As of the date of testing this girl was 9 years and 1 month of age, making her five months younger than the average of her grade at the time of testing. In this respect, she is similar to the previous case. However, she differs radically in that her Deviation IQ on the Otis-Lennon Mental Ability Test is 128 and her Otis-Lennon percentile rank by age is .96 and by grade is .93.

Turning now to the data tabulated at the bottom of the Individual Profile Chart and Personal Data Sheet and considering first her fall performance, we see that she has made use of the Omits option in every one of the five achievement tests as well as the Otis-Lennon, where she had a Right score of 57, a Wrong score of 15 and Omitted 8 items.

Her R/A ratio is high for Word Meaning and Paragraph Meaning, as it is for the Otis. It also is high for Arithmetic Concepts and only slightly lower for Arithmetic Applications, but in both instances her R/A ratio is higher in the fall than it is in the spring. This is a reversal of the situation found for the group as a whole.

Her R/A ratios for Arithmetic Computation are low, quite unsatisfactory as a matter of fact, with values only of .32 and .34 for fall and spring respectively. The reasons for this become perfectly evident when you study the data on the Item Analysis Data sheet.

This girl's gains from fall to spring are notable in Word Meaning and Paragraph Meaning, but it is evident that something has gone very badly awry in the arithmetic field.

Considering first Arithmetic Computation, she makes a gain from fall to spring of only 1 point, going from 9 to 10, and both of these scores are very near the average guessing level.

In Arithmetic Concepts her scores are higher, but her gain in score is only from 19 to 21, or 2 points of score compared to about 4 points average gain for the random sample.

Her scores are more reasonable in Concepts than they are in the other fields of arithmetic, but the bloom is taken off the blossom to some extent by noting that she begins a guessing pattern almost from the very beginning. She has two "WW" responses in the first ten easy items; she has two "RW" responses in the first fifteen items, another for item #22; she attempts all items in the spring but misses six out of the last fifteen items.

The items she answered "RW" are particularly serious, as such responses are almost a precise indictment of the response as being a guessed response.

Arithmetic Applications in many ways is the most peculiar of all the tests. Only in the first block of five items does she demonstrate knowledge that you can depend on. In the second block, three of her fall responses which were correct become incorrect in the spring ("RW"); in the third block, three responses were Right in the fall and Wrong in the spring; before the end of the test she has reversed two more responses that were originally Right to Wrong in the spring, making a total of eight "RW" responses.

Her last three groups are all suspect re guessing. Finally, her score of 17 in the fall drops to a score of 15 in the spring, implying no gain at all during the seven months' period; in fact, a loss of 2 points as compared to a gain for the total group of 4 points.

Her Predicted Scores coincide fairly well with her earned scores for all tests both fall and spring.

The stanine profile of her performance, remembering again that these are separately computed stanines for fall and spring, would suggest that she has moved along pretty much in synchronization from fall testing to spring, there being no greater difference between stanines than 1 point. A drop of 1 point is to be found in each of the three Arithmetic Tests, while a gain of 1 point is to be found in Paragraph Meaning. In Word Meaning, her stanine is identical.

With her Otis-Lennon score and DIQ, she should have greatly exceeded state average performance. Hence, she definitely has an arithmetic problem. Lack of mastery of fundamentals is the best guess because this is a familiar pattern for very bright children who often are lax in rote learning.

It has been repeatedly stated throughout this report that Word Meaning and Paragraph Meaning are subjects in which status on a standardized test depends almost as much on what happens outside of school as on school-learned knowledges and skills.

Answer Sheet Study - II

Of the three Arithmetic Tests, Concepts draws most heavily on the reasoning-type factors to be found in Otis-Lennon, and this is the one arithmetic area where she is above the statewide random sample in the spring.

Conclusion

This child is not working up to capacity in any test and clearly has what would amount to a specific disability in arithmetic that very often characterizes the very bright child to whom over-learning of basics (e.g., 100 addition and subtraction facts and multiplication tables) as a result of drill and constant repetition is boring and distasteful.

If this writer were dealing with this particular child, his first step would be to investigate the arithmetic area more closely by identifying the pattern of answering the items for each of the three Arithmetic Tests to discover the kinds of mistakes the child is making, and specifically to decide whether or not these errors were largely due to probable lack of mastery of the addition-subtraction facts and of the multiplication tables, as suggested.

PART III

COMPARISON OF TITLE I CASES WITH THE RANDOM SAMPLE ON ALL ESSENTIAL VARIABLES

INTRODUCTION

From the beginning, we have emphasized the fact that we have two populations being treated identically so far as testing is concerned; including the time of the administration of the tests, the conditions under which they are administered, and all other similar variables.

This section of the study is similarly organized. The analyses of data follow the general format found in Part II. This study is not to be confused with an earlier report which also involved Title I children for the whole state as well as the random sample of the state. 1/

Our first conclusion must be that Title I children are, indeed, different from the total population (or from the random sample, already shown to be representative), and this difference runs through every test administered and all subsequent analyses.

It would become rather boring and less productive to make a routine comparison exactly the same as was done for the random sample, so we will concentrate on differences.

It is essential that a sufficient amount of detailed comparison should be built into this report to convince one that essential generalizations change greatly when one moves from a population such as the random sample to the Title I group - other than the fact that the Title I group performance drops on the score scales.

The test in question may be a mental ability test with a Deviation IQ, or it may be Paragraph Meaning, or Arithmetic Computation, or Science. In every case, there is a drop from random sample to Title I.

In no case does the average of the Title I group reach or exceed that of the random sample, but in every case some of the children in the Title I group do reach or exceed the average score of children in the random sample. In other words, there are overlapping distributions.

No additional research was needed to reach this conclusion; it is inevitable because of the substantial variability of children's ability and the spread of entrance age over at least a year's span.

Some of the very bright children in the state have been included among the Title I children in this study, for reasons which cannot now be ascertained because they are local and expedient in nature.

We can only assume that (in part, at least) the reason arises from the somewhat unrealistic basis by which the law provides for the selection of these individuals. Some of the other more evidential reasons will be discussed later.

The Original Title I Report

It is pertinent to remind the reader that the basic score comparisons of the Title I population and the random sample, as well as the state as a whole, were done in great detail in the first report entitled: "A Description and Evaluation of the Statewide Testing Program in New Hampshire in 1968-69 and 1969-70."

Much of the data in the first part of this section comes directly out of the earlier report, and it is highly recommended to anyone who is making a careful study of this report that he obtain the earlier report first to provide the necessary background.

Section VI of the original report is specifically concerned with the comparison of the random sample with the total state population versus the Title I group.

To save the time and bother of consulting the earlier report, or in some cases its unavailability to the reader, we feel impelled to repeat here some of the essential findings:

1. The Title I sample available for study cannot in any way be considered a random sample from the entire state - nor even of the group which normally would be considered eligible for Title I assistance by strict adherence to the law.

Some of the larger cities in the state chose to go their own way so far as evaluation was concerned, and there is nothing in the national law to prevent their doing so.

Hence, to some extent our Title I population must be considered a biased sample of all Title I cases in the state. In

---

1/ "A Description and Evaluation of the Statewide Testing Program in New Hampshire in 1968-69 and 1969-70 Under the Sponsorship of Title I and the Significance of the Data Obtained for Evaluation With This Activity." Prepared by the Test Service and Advisement Center. 1971.

general, the bias would not be in the direction of increased ability level of the Title I children tested in our group. It might even have decreased it. Nobody knows for sure.

We can be sure it was the Title I sample included in this federally funded program in New Hampshire. This is important because it bears upon the extent to which generalizations can be made to other Title I programs in other states.

2. The composition of the random sample group had approximately equal numbers of boys and girls; on the other hand, the Title I group is disproportionately boys, having 61% versus 39% girls.

One could consider this disproportionate maleness to be a local bias if it did not happen so frequently in so many studies of the disadvantaged or handicapped, delinquent, or poor-achieving child in school.

(This writer did numerous studies, for example, of delinquency and emotional instability in the schools of Pinellas County, Florida, and found repeatedly that about two-thirds of these cases were boys. The writer also was in direct charge of the corrective reading program in the county, and here, again, about two-thirds of the children under instruction in the corrective reading program were boys.)

The literature is full of comparisons of this sort; therefore, one must assume that whatever the basis is for choosing the children for studies of this sort it has very generally resulted in about the same disproportionate number of boys compared to girls.

3. The Title I sample is older than the random sample.

This follows from the arbitrary and unreasonable entrance requirements held to almost uniformly throughout the state - and most other states, for that matter. Minimum age for entrance into grade one in this state varies, but generally children have to be 6 years old not later than December 1.

The difficulties found by Title I children quickly show up and result in retardation unless the school system has an ungraded primary system.

The Title I boys averaged 9 years and 11 months of age; girls averaged 9 years and 8 months of age - an interesting phenomenon. The total population of Title I averaged 9 years and 10 months of age.

The median age at the beginning of the fourth grade testing in October for the state as a whole was about 9 years and 6 months.

By contrast, the random sample of children chosen for our study was slightly younger and brighter than average, being 9 years and 4 months of age; but this was a factor over which the investigator had no control.

The fact that this sample was tested in the spring for our convenience resulted in a substantial reduction in the number of cases identified by computer to be included in this study, as discussed in the original Title I report.1/

Comparing totals only, it appears, then, that there is a 6 months age difference between the random sample (complete cases only) and the Title I group.

4. 92% of the Title I children fell at or below the average random sample Deviation IQ of 102.

The Otis-Lennon Mental Ability Test: Elementary II Battery: Form J was administered statewide at the beginning of the test program (Fall 1969). The Title I boys earned an average Deviation IQ of 85; the girls, 88; and the total was 86. The average for the entire statewide population in Grade 4 was 101.

At this point one must ask oneself if it was the intent of the lawmakers who framed Title I to provide a program for slow-learning children - which, in effect, it did.

The answer is emphatically, "No!"; it was to provide a special opportunity for children coming from disadvantaged backgrounds.

We have no way of knowing that the average mental ability (as measured) of the parents of these children was lower than for the population as a whole, or whether the lower IQ's of the children in this program were due to the disadvantages under which they lived.

It's specious to say that a child does better if he's under stimulating circumstances at home (and/or in his general environment) than he does if he's in a restricted and impoverished environment.

Neither the disadvantaged alone nor even the most fortunate people in the

---

1/ Ibid.

state in regard to environment have any monopoly on brightness. Very many of our ablest people, in the history of this country especially, have come from homes of great poverty and hardship with very few opportunities to "make something of themselves" except as they went out and found these opportunities on their own initiative.

As new technologies develop in this technological world, they're going to develop because some people with creative ideas, regardless of their backgrounds, bring to them the dedication it takes to stick to something until the job is done.

What Mental Ability Tests Are

Mental ability tests are nothing but a series of tests to roughly sort out and bring some order to the hierarchy of ability. They consist of real-life problems, generally not school-oriented, which are stated in verbal terms but which require for their solution a variety of skills.

Sometimes they involve knowledge of vocabulary. Sometimes they involve solving problems seemingly related to mathematics and physics, in the type of thinking involved, but stated in simple and untechnical terms.

All of the problems involved in any good mental ability test are oriented specifically to the whole environment as the source of knowledge. The greater harm, however, lies in employing a test of this sort with the disadvantaged child who, because of his meager background, is unable to cope on equal terms with someone no more alert or of no greater mental ability than he.

The tests reflect, but do not measure, the magnitude of the "disadvantage" as regards school accomplishement.

We need to be sure, for example, to choose a test such that nothing in the test-taking experience adds to his difficulties. Practice sheets, careful oral instruction, time for questions before testing, etc., all can help.

Test-taking skill can and should be taught as a prerequisite of actual test administration. However, by the beginning of the fourth grade few pupils will not have been subjected to objective testing in this state.

Moreover, if one divests himself of the idea that these tests measure native intelligence or something that cannot be changed by enriching and expanding horizons, most of our "hangups" disappear. For all practical purposes, tests of this sort reflect what a child is able to do at a particular moment, but not necessarily what he will be able to do if he is given proper stimulation.

The sad part is not the instability of mental ability measures, but their consistency over a wide span of years.

Perhaps of greatest importance of all is the fact that the general mental ability test is the one test that correlates most highly with almost any other measured school-learned skill. This is not only true of reading and vocabulary, which are themselves saturated with language, but it is equally true of mathematics - and particularly so of concepts and applications.

It certainly is true of certain aspects of science and social studies testing also, especially in the middle and upper grades - and especially in the more modern textbooks where there is a diminution of emphasis on knowledge of certain facts about history, social studies, or science in favor of the development of skills in adapting to new facts (which are developing all the time) and in the development of ability to find and assess information of relevance to some problem that needs to be solved at the moment.

Finally, it is of greatest importance to emphasize in this study that the Title I children studied were not chosen on the basis of any mental ability measure or even on the basis of a systematic achievement test program. All of these came after the fact, so to speak.

Children had already been selected and allocated to Title I projects before the opening of school, so it remained to administer the tests to these children, along with all of the other children in the grades involved (2, 4, 6, and 8 in the years indicated), in October and to re-test the Title I children and the random sample in the following late April and/or early May.

Differentiation of Achievement Tests

Let us now take a brief look at the achievement tests - namely, the Stanford Achievement Test: Intermediate I Battery: Form X - and try to make a judgment as fairly as possible as to the extent to which the content of these tests is biased in favor of one socioeconomic group as compared to another.

In doing this, it has to be remembered that there is no single very prominent low socioeconomic group in the State of New

Hampshire (as in the South or in our national metropolitan areas). There may be something of a bilingual problem in the northernmost counties and in some of the southern cities, but the proportion of bilingual children is very small.

Looking first at the Word Meaning Test, it is very important to remember that these words came from materials found appropriate for the grade level in terms of the vocabulary widely used in textbooks in this grade. This was true at the time the words were selected to be tested and the items written, but the process of item analysis eliminated words which were non-functioning for the total group; i.e., words too simple or too complex.

There were, indeed, hard items and easy items left in the test; otherwise, the least able and the most able children in the group tested would not have been able to make an acceptable unbiased score.

Incidentally, as we get into this study we must conclude that the difference between a survey test, such as Stanford, and a test intended from the beginning to measure the before-after performance at a single grade in a single state or community is very great; indeed, it is much greater than any of us realized, perhaps, until this study was carried out.

Paragraph Meaning, as contrasted to vocabulary development alone, does include the development of certain specifically taught skills - such as the ability to make a phonic attack on new words; that is, to derive the meaning of words that are new to the child in their written form - at least in the context of this test.

It only happens very rarely, and particularly with children who are low in general mental ability, that a word a child might be expected to learn to read is not already known to him when it is spoken. The child's spoken vocabulary tremendously exceeds his written or reading vocabulary at the time he goes to school, and probably through the lower elementary grades. For many people, this remains to be true through their whole lives.

This is the essence of reading instruction in the lower grades; namely, learning the written symbol that stands for a particular word we already know when spoken. Later, the process may be reversed; we may learn to speak and write words encountered first in reading! This, however, occurs only in the higher grades among children already rated as good readers.

There have been, over the last couple of decades, violent controversies as to whether the look-say method (or the whole method, as it is sometimes called) is better than the phonic method.

There are arguments to be made for both approaches, but probably the most unbiased and uncommitted study done in this area seems to indicate that method makes relatively little difference, provided the teacher adapts his or her instruction to the need of the individual child. 1/

In Arithmetic Computation, as contrasted to most other school subjects, there are certain basic knowledges and skills which have to be mastered, and a lack of mastery of these skills constitutes a continuing handicap throughout one's life.

For example, if a child does not know his 100 addition and subtraction facts and eventually his multiplication tables, he will be handicapped constantly in doing other kinds of arithmetic.

He may be able to think his way through certain abstractions in advanced mathematics, which really involves little manipulation of numbers but rather encompasses constellations of ideas concerning the relationships between quantitative ideas.

Is all this repetitive? Well, perhaps so. It will stand repeating. Some people may read only Parts III and IV!

1/ Chall, Jeanne. Learning To Read: The Great Debate. New York: McGraw-Hill, 1967.

THE TITLE I DATA COMPARED WITH RANDOM SAMPLE

Enough has been said in the previous paragraphs to lead us directly into a comparison of the data for Title I that is strictly comparable in its nature to the data previously presented for the random sample.

In the random sample section of this report (Part II), we gave as histograms the actual distributions of Word Meaning in the fall and spring and also Arithmetic Computation in the fall and spring. We will do the same thing for Title I.

A test which is used at the beginning and again at the end of instruction, which was true in this case, needs to be on the hard side at the beginning in order to give the pupils an opportunity for the maximum amount of learning during the period of instruction.

Coupled with this, of course, is the corollary that the material that is not known at the start is material to which the child will be exposed during the course of the year's instruction.

Therefore we have to be very careful to be sure that the test does, indeed, measure what the teacher intends to teach - not so much specific-item-by-specific-item as in a broad, general way; e.g., not so much the meaning of "attachment," as the broad skills in method attack which will help the child learn this word.

The Word Meaning Test is broadly based, especially as to type; but not with the idea that these identical problems define the curriculum. The child's eventual vocabulary is much larger than any curriculum in word meaning.

The test sample is so small a sample of the total vocabulary that neither this test nor any other group of commonly used words will, of a certainty, be found in the local curriculum.

Word Meaning Content Related to the Title I Score Distribution

These words represent a cross section or random sample of the kinds of words children at grade four are likely to encounter, plus a good saturation of words that the children should have been exposed to at grade three and some harder ones to give "top" to the test.

In Figure III-1, we give the Word Meaning distribution for the fall. If one looks back at the similar distribution for the random sample (Figure II-1), it is easy to see that this test was much harder for Title I than it was for the random sample; and yet the Title I fall distribution does have cases earning scores as high as 32 out of a 38-item test.

The mean of 9.13 is substantially lower than the random sample mean of 15.92, and the random sample is considerably more variable - as indicated by the comparative standard deviations.

The point, however, is that the group selected for Title I does distribute itself across the continuum of vocabulary as measured by Stanford.

The important thing to note regarding the Title I fall distribution (Figure III-1) is not so much the piling up at the lower end, but the fact that the median (and modal) value in this distribution is only a score of 8!

The average chance score on this distribution of four-choice items would be 9.5 questions answered correctly out of the 38 items, which is the number of items included in this test. The mean score of 9.13, as a matter of fact, is slightly below the chance level (9.5).

However, when the children were retested in the spring, the mean had moved up to 13.2 (Figure III-2) and, although the gain of 4 words (or points of score) between October and May is certainly not anything to be gleeful about, the test at this point does not look much different than the kind of distribution we very often get with a survey-type standardized achievement test with similar groups.

Obviously, all those who earned scores below 9 or 10 did not do so by chance alone; the problem is (and always has been): How many correct responses were obtained by guessing?

We do not give nearly enough emphasis to the fact that in tests of this sort there are large numbers of children who are clearly working so far below their grade level that they simply do not have the opportunity to progress very far above the guessing level during the relatively short instructional period of time involved (seven months).

If vocabulary building in the local situation is not specifically a goal of instruction but is left to incidental learning in connection with all instruction, no standardized vocabulary test is curriculum valid (in the strictest sense) at the local level. Not even a locally-made test would be valid,

Raw Score	Cum. %	Sta-nine	Fre-quency	
32	99	9	1	*
31	99	9	0	
30	99	9	1	*
29	99	9	0	
28	99	9	0	
27	99	9	0	
26	99	9	1	*
25	99	9	1	*
24	99	9	1	*
23	99	9	3	***
22	98	9	4	****
21	97	9	2	**
20	97	9	2	**
19	96	8	6	******
18	95	8	12	************
17	92	8	4	****
16	91	8	9	*********
15	89	7	5	*****
14	88	7	15	***************
13	84	7	18	******************
12	80	6	29	*****************************
11	74	6	23	***********************
10	68	6	34	**********************************
9	60	5	39	***************************************
8	51	5	43	***
7	41	4	38	**************************************
6	33	4	40	**
5	23	3	30	******************************
4	16	3	28	****************************
3	10	2	25	*************************
2	4	1	9	*********
1	2	1	9	*********
			432	

FIGURE III-1
Frequency Distribution, Cumulative Percent Distribution, and Stanines Plus Histogram Showing Shape of Raw Score Distribution Graphically

TITLE I - WORD MEANING - FALL 1969*

Mean - 9.13

St.Dev. - 4.98

* Each * = one case

Answer Sheet Study - III

```
 Raw   Cum.  Sta-  Fre-
Score   %    nine  quency

 34    99     9      1    *
 33    99     9      1    *
 32    99     9      2    **
 31    99     9      1    *
 30    99     9      1    *
 29    99     9      5    *****
 28    97     9      0
 27    97     9      2    **
 26    97     9      3    ***
 25    96     8      4    ****
 2[illegible]    95     8      5    *****
 23    94     8      5    *****
 22    93     8     12    ************
 21    90     7     11    ***********
 20    88     7     14    **************
 19    85     7      8    ********
 18    83     7     21    *********************
 17    78     6     21    *********************
 16    73     6     33    *********************************
 15    65     6     21    *********************
 14    61     5     26    **************************
 13    55     5     26    **************************
 12    49     5     16    ****************
 11    45     5     23    ***********************
 10    40     4     30    ******************************
  9    33     4     40    ****************************************
  8    24     3     25    *************************
  7    18     3     19    *******************
  6    13     2     25    *************************
  5     8     2     15    ***************
  4     4     1     11    ***********
  3     2     1      3    ***
  2     1     1      3    ***
  1     1     1      1    *
                   434
```

FIGURE III-2
Frequency Distribution, Cumulative Percent Distribution, and Stanines Plus Histogram Showing Shape of Raw Score Distribution Graphically

TITLE I - WORD MEANING - SPRING 1970*

Mean - 13.21 St.Dev. - 6.11

* Each * = one case

either, unless a careful study was made of the words taught during the school year.

The closest approach to a valid test at the local level would be the word list accompanying the reading series, plus (as a matter of personal opinion) the words taught as part of spelling instruction.

It is much more important to measure the children at the low end of the achievement scale than it is to measure those at the top, since those at the top will exceed by substantial margins the average vocabulary performance of children at their grade, due to general environmental factors as well as instruction.

The Average Chance Score

We also see that 261 children in this group of 432 Title I children taking the Word Meaning Test in the fall (Figure III-1) achieved scores of 9 or lower, which means that they scored essentially at the chance level. In other words, if these children had simply marked the paper without ever looking at it, they would have a 50/50 chance of getting as high a score as they earned.

The average chance score depends on the number of alternative choices provided and the number of items. It is a fraction with "1" as the numerator and the number of alternatives as the denominator times "n" items. Thus, for a four-choice test it is ¼n, where "n" is the number of items in the test.

Arithmetic Computation Distribution Characteristics

Turning now to the Arithmetic Computation Test, it is evident that this test is also too hard. It is too hard even for the random sample. It is a 39-item test, but the highest score obtained by anyone is only 29 in the fall random sample group.

Arithmetic Computation is a five-choice item test; and therefore by chance, on the average, marking the answer sheet without reference to the test would give an individual a score of 1/5 of the total number of items (39), or 8 items right in round numbers.

(Eight is the average of a normal distribution of errors for 39 five-choice items but the standard deviation of this distribution, which cannot be exactly obtained by any simple method, is probably about 4 or 5 points.)

In the random sample in the fall program, 26% of the children had scores that could have been obtained as frequently as not by chance, assuming all 39 items had been marked. This dropped to 8% in the spring.

By and large a very substantial majority of children answer the questions they know and omit many of the remaining items; these they do not mark by chance, obviously. This is as it should be.

This study is most revealing in showing that the proportion of those who do use chance marking is substantially greater than we had suspected it might be. We must ask ourselves very seriously if this is a tolerable situation.

But what about the wrong responses? What proportion of wrongs to rights is acceptable? In a work-sample type item, where there is little or no chance of guessing, the proportion would be zero!

The situation naturally is worse in the case of Title I, with 41% of the children achieving scores at the average chance level or below in the fall. (See Figure III-3) This means, of course, that there will be a roughly equal number of others who most likely have earned higher scores by chance; they were among the lucky ones in their choice of correct answers, if you take their point of view.

All this theory applies only when all the items have been marked. For a child who attempts 29 of a 39-item test, 29 (not 39) is the effective test length for that child and the chance situation is changed. This is the fallacy of the traditional correction for chance.

Remembering now that we have concluded prior to this that a difficult test is desirable at the beginning of the instructional period, we note that the mean for the random sample was 11.46 in the fall but that this jumped to 18.34 in the spring, or a total of about 7 points during the course of the seven-months period between fall and spring testing.

For the Title I group, the gain is only 4½ points, but (in view of the fact that this is a much less able group) this is a notable gain by comparison. (See Figure III-4.) There still remains about 17% of the group, even in the spring, who are at the average chance level or below, all other previously stated conditions applying.

```
 Raw   Cum.  Sta-  Fre-
Score   %    nine  quency

 21     99    9      2   **
 20     99    9      4   ****
 19     99    9      6   ******
 18     97    9      9   *********
 17     95    8     14   **************
 16     92    8     10   **********
 15     90    7     18   ******************
 14     85    7     20   ********************
 13     81    7     26   **************************
 12     75    6     32   ********************************
 11     67    6     39   ***************************************
 10     58    5     33   *********************************
  9     41    5     42   ******************************************
  8     41    4     43   *******************************************
  7     31    4     41   *****************************************
  6     22    3     47   ***********************************************
  5     11    2     19   *******************
  4      6    2     15   ***************
  3      3    1      9   *********
  2      1    1      2   **
  1      1    1      2   **
                   433
```

FIGURE III-3
Frequency Distribution, Cumulative Percent Distribution, and Stanines
Plus Histogram Showing Shape of Raw Score Distribution Graphically

TITLE I - ARITHMETIC COMPUTATION - FALL 1969*

Mean - 9.94 St.Dev. - 4.03

* Each * = one case

Raw Score	Cum. %	Sta-nine	Fre-quency	
37	99	9	1	*
36	99	9	1	*
35	99	9	0	
34	99	9	1	*
33	99	9	0	
32	99	9	1	*
31	99	9	1	*
30	99	9	3	***
29	99	9	6	******
28	97	9	3	***
27	96	8	6	******
26	95	8	5	*****
25	94	8	4	****
24	93	8	8	********
23	91	8	7	*******
22	89	7	10	**********
21	87	7	7	*******
20	85	7	20	********************
19	81	7	20	********************
18	76	6	18	******************
17	72	6	22	**********************
16	67	6	21	*********************
15	62	5	29	*****************************
14	55	5	32	********************************
13	48	5	28	****************************
12	41	4	27	***************************
11	35	4	26	**************************
10	29	4	35	***********************************
9	21	3	19	*******************
8	17	3	14	**************
7	14	3	26	**************************
6	8	2	14	**************
5	4	1	8	********
4	3	1	7	*******
3	1	1	1	*
2	1	1	1	*
1	1	1	2	**
			434	

FIGURE III-4
Frequency Distribution, Cumulative Percent Distribution, and Stanines
Plus Histogram Showing Shape of Raw Score Distribution Graphically

TITLE I - ARITHMETIC COMPUTATION - SPRING 1970*
Mean - 14.46 St.Dev. - 6.28

* Each * = one case

Ambiguity in the Selective Process for Title I Cases

The selection of children for Title I projects did not clearly differentiate between those with remedial or correctable defects, especially in reading, as compared to those who were simply slow in developing their verbal learning ability. This is especially significant in the lower grades where the foundations of reading skills are being established.

Lack of recognition of this duality shows itself particularly at the fourth grade level, when one begins to move from the teacher-oriented presentation of material (in lecture form or in demonstration) to the textbook as the main source of information; i.e., new knowledge.

If a child has failed to learn to read because he is slow in maturing and yet he finds himself at the fourth grade level because of chronological age promotion, a difficult situation indeed is created for both the teacher and the pupil.

If the child's problem is a correctable reading difficulty with good potential indicated by independent evidence (not the least of which is the teacher's own judgment), the opportunities for positive change are much greater - and consequently the need for careful analysis and diagnosis in the reading area is also greater.

There is little in the evidence available to us to differentiate between these two kinds of cases.

In the remedial reading program in Pinellas County, Florida, for which the writer was largely responsible, the oral vocabulary was thought to establish the upper limits of reading ability through the third grade.

Standardized tests (like Stanford: Intermediate I: Paragraph Meaning) were given twice. The first time they were given in the standard manner, and this was followed by the administration of a second form where the teacher read out loud while the children read silently.

Many children, even among those considered to be at-grade-for-age or above, were discovered to be potentially much better readers than was shown by test results. Many others whose ability was average or less than average were identified as needing the assistance of specialists because of the existence of correctible defects requiring both time and techniques not available to the classroom teacher.

The analytical and diagnostic program designed to determine the nature of a child's difficulty is something that is beyond the scope of the average classroom teacher's training, and often specialized equipment is required.

In Title I, even more than in the population as a whole, it becomes very essential to determine whether a child's performance (especially if it is well below grade level) is low because he is just naturally a slow learner or because he has a correctible defect in reading.

In this respect, the test reflects the classroom. If he is a child who is average in verbal abilities or slightly below average, the likelihood is much greater that he has not benefited as much from the instruction in the first three grades as could be expected.

The same type of observation described above should be carried out for the subsequent tests in this battery. In order to make this simpler, Table III-1 has been prepared giving the mean, the median, and the standard deviations for each of the five tests studied both fall and spring.

Table III-1
RAW SCORE Means, Medians, and Standard Deviations

TITLE I

Tests	No. of Items	Mean		Median		Stand. Dev.	
		Fall	Spring	Fall	Spring	Fall	Spring
Word Meaning	38	9.1	13.2	8.0	12.0	5.0	6.1
Paragraph Meaning	60	16.3	20.9	15.5	19.0	6.0	7.9
Arithmetic Computation	39	9.9	14.5	9.5	13.0	4.0	6.3
Arithmetic Concepts	33	9.4	11.7	8.0	10.0	4.4	5.2
Arithmetic Applications	32	8.9	10.9	8.0	9.5	4.3	5.5

As regards the histograms themselves, the single generalization that one must inevitably reach is that any test in the Stanford: Intermediate I Battery will reveal an enormous spread of ability even in Title I.

Some Aspects of Arithmetic Applications

Let us look at a test which probably depends very largely on sheer mental ability independent of any total environmental effects; namely, Arithmetic Applications (commonly called Problem Solving) - in which the manipulation of numbers is now applied in a verbal setting to obtain the solution of simple problems.

Here, again, we see a skewed distribution in the fall; that is, one with a tail running off toward the high scores and the distribution tending to be piled up toward the beginning of the test. (Figure III-5.)

We must consider the fact that this is a very short test (having only 33 items), but these are five-choice items; therefore, the average chance score would be 1/5 of 33, or about 6 or 7 points.

In the fall, 31% of the 434 children had scores of 6 or lower. The mean score is only 8.9, and the standard deviation is 4.25. The median score is about 8.

When we look at the distributions for this test in the spring, seven months later (Figure III-6), we find the mean of the Title I group has increased from 8.9 to 10.9, or only 2 points. This gain must be evaluated in terms of a comparison with the gain made by the state as a whole and also by the national group.

The national gain over a seven-month period, as near as can be estimated by interpolation from the norms given in the accessory materials, is about 4½ points.

For the New Hampshire random sample, the gain was 3.3. Thus the Title I group gained less in proportion than the random sample, but about as one would expect or perhaps even a little better.

Obviously a gain of 2 points, or even of 3.3 points, over a period of seven months is nothing to brag about in terms of the quality of an evaluation of the effectiveness of instruction across such a period of time, involving as it does the major part of the instructional school year.

Even after seven months of instruction, less than 25% of the children answer more than 14 items correctly. The last half of the test, for all practical purposes, was quite ineffective - being too hard for even the higher ability levels of children in this group.

Is this, then, a good test of arithmetic problem solving ability to use with this group; specifically, to find out how much learning takes place from fall to spring for Title I children?

It cannot be said to be a good test, even for the total group, because it is a potpourri of problem solving items; i.e., an average-type curriculum based on many textbooks. Norms were obtained by administering the final test in the spring to very large populations of children all over the country.

Here, we are not concerned with norms. We are, however, vitally concerned with the local curriculum and all other factors affecting success - especially in this, the level of reading ability.

We are looking at distributions of raw scores for our own Title I population alone, and it only consists of 434 children out of the 10,000+ who were tested at this grade.

However, these were the children selected for special projects under the provisions of Title I because of need. Therefore for them, if not for the total group, a test more ideally suited to the local curriculum is quite essential.

In fairness to Stanford, it should be said that this particular edition was published at a time when modern mathematics was just beginning to appear on the horizon, and for some people this test was particularly hard in math simply because it did try to at least sample some modern math concepts, although it could not really get very deeply involved in such content.

Through all our studies, we have emphasized the large amount of guessing taking place and have suggested corrective steps which would result in an improved situation. If no guessing were involved, it would depress the means - and the distributions would be even more truncated than they are.

(See Figures III-7 - 10 for the distributions of raw scores for Paragraph Meaning and Arithmetic Concepts, Title I, fall and spring.)

Fewer children would get high scores simply because random guessing must inevitably add to the scores for at least 50% of the participants in the testing program who guessed when the answer was not known.

When all do not guess equally, even a point or two of score obtained sheerly by guessing tends to dilute, weaken, and to some extent destroy the meaningfulness of the test data for the entire group, especially for instructional purposes.

The significance of all this will become much clearer as we proceed with our analysis.

Raw Score	Cum. %	Stanine	Frequency	
25	99	9	1	*
24	99	9	2	**
23	99	9	1	*
22	99	9	0	
21	99	9	1	*
20	99	9	2	*
19	98	9	3	***
18	98	9	5	*****
17	97	8	8	********
16	95	8	12	************
15	92	8	12	************
14	89	7	18	******************
13	85	7	18	******************
12	81	7	25	*************************
11	75	6	22	**********************
10	70	6	30	******************************
9	63	5	47	***
8	52	5	47	***
7	41	4	44	**
6	31	4	39	***************************************
5	22	3	38	**************************************
4	14	2	28	****************************
3	7	2	16	****************
2	3	1	13	*************
1	1	1	2	**
			434	

FIGURE III-5
Frequency Distribution, Cumulative Percent Distribution, and Stanines Plus Histogram Showing Shape of Raw Score Distribution Graphically

TITLE I - ARITHMETIC APPLICATIONS - FALL 1969

Mean - 8.89

St.Dev. - 4.25

* Each * = one case

```
 Raw   Cum.  Sta-   Fre-
Score   %    nine  quency

 32    99     9      1   *
 31    99     9      0
 30    99     9      0
 29    99     9      1   *
 28    99     9      0
 27    99     9      0
 26    99     9      3   ***
 25    99     9      1   *
 24    99     9      6   ******
 23    97     9      4   ****
 22    96     8      7   *******
 21    95     8      6   ******
 20    93     8     10   **********
 19    91     8      5   *****
 18    90     7      9   *********
 17    88     7     12   ************
 16    85     7     22   **********************
 15    80     7     15   ***************
 14    76     7     17   *****************
 13    73     6     34   **********************************
 12    65     6     22   **********************
 11    60     5     27   ***************************
 10    53     5     27   ***************************
  9    47     5     32   ********************************
  8    40     4     33   *********************************
  7    32     4     47   ***********************************************
  6    21     3     23   ***********************
  5    16     3     28   ****************************
  4     9     2     19   *******************
  3     5     1     10   **********
  2     3     1      6   ******
  1     1     1      6   ******
                    ---
                    433
```

FIGURE III-6
Frequency Distribution, Cumulative Percent Distribution, and Stanines
Plus Histogram Showing Shape of Raw Score Distribution Graphically

TITLE I - ARITHMETIC APPLICATIONS - SPRING 1970

Mean - 10.88 St.Dev. - 5.53

* Each * = one case

Raw Score	Cum. %	Sta-nine	Fre-quency	
41	99	9	1	*
40	99	9	0	
39	99	9	0	
38	99	9	1	*
37	99	9	0	
36	99	9	1	*
35	99	9	2	**
34	99	9	1	*
33	99	9	1	*
32	98	9	1	*
31	98	9	3	***
30	97	9	2	**
29	97	9	3	***
28	96	9	2	**
27	96	8	3	***
26	95	8	5	*****
25	94	8	8	********
24	92	8	9	*********
23	90	7	14	**************
22	87	7	15	***************
21	83	7	23	***********************
20	78	6	16	****************
19	74	6	20	********************
18	70	6	38	**************************************
17	61	5	26	**************************
16	55	5	38	**************************************
15	46	5	40	**
14	37	4	30	******************************
13	30	4	25	*************************
12	24	3	28	****************************
11	18	3	15	***************
10	15	3	13	*************
9	12	2	15	***************
8	8	2	8	********
7	6	2	6	******
6	5	1	10	**********
5	3	1	4	****
4	2	1	3	***
3	1	1	3	***
2	1	1	1	*
			434	

FIGURE III-7
Frequency Distribution, Cumulative Percent Distribution, and Stanines Plus Histogram Showing Shape of Raw Score Distribution Graphically

TITLE I - PARAGRAPH MEANING - FALL 1969

Mean - 16.35 St.Dev. - 5.97

Raw Score	Cum. %	Sta-nine	Fre-quency	
54	99	9	1	*
53	99	9	0	
52	99	9	0	
51	99	9	0	
50	99	9	0	
49	99	9	1	*
48	99	9	0	
47	99	9	0	
46	99	9	0	
45	99	9	1	*
44	99	9	3	***
43	99	9	1	*
42	98	9	1	*
41	98	9	2	**
40	98	9	2	**
39	97	8	1	*
38	97	8	2	**
37	97	8	2	**
36	96	8	2	**
35	96	8	4	****
34	95	8	8	********
33	93	8	3	***
32	92	8	11	***********
31	90	7	9	*********
30	88	7	12	************
29	85	7	11	***********
28	82	7	7	*******
27	81	7	12	************
26	78	6	10	**********
25	76	6	12	************
24	73	6	19	*******************
23	68	6	18	******************
22	64	6	20	********************
21	60	5	15	***************
20	56	5	25	*************************
19	50	5	23	***********************
18	45	5	32	********************************
17	38	4	24	************************
16	32	4	31	*******************************
15	25	4	18	******************
14	21	3	18	******************
13	17	3	18	******************
12	13	2	23	***********************
11	7	2	8	********
10	6	2	7	*******
9	4	1	7	*******
8	2	1	4	****
7	1	1	2	**
6	1	1	2	**
5	1	1	1	*
4	1	1	0	
3	1	1	0	
2	1	1	0	
1	1	1	1	*
			434	

FIGURE III-8
Frequency Distribution, Cumulative Percent Distribution, and Stanines Plus Histogram Showing Shape of Raw Score Distribution Graphically

TITLE I - PARAGRAPH MEANING - SPRING 1970

Mean - 20.85 St.Dev. - 7.93

Raw Score	Cum. %	Stanine	Frequency	
27	99	9	1	*
26	99	9	1	*
25	99	9	1	*
24	99	9	0	
23	99	9	1	*
22	99	9	2	**
21	99	9	0	
20	99	9	6	******
19	97	9	6	******
18	96	8	8	********
17	94	8	7	*******
16	92	8	9	*********
15	90	7	15	***************
14	87	7	14	**************
13	84	7	17	*****************
12	80	6	24	************************
11	74	6	34	**********************************
10	67	6	40	**
9	57	5	42	**
8	48	5	41	***
7	38	4	47	***
6	28	3	44	**
5	17	3	30	******************************
4	11	2	20	********************
3	6	2	17	*****************
2	2	1	6	******
1	1	1	3	***
			436	

FIGURE III-9
Frequency Distribution, Cumulative Percent Distribution, and Stanines
Plus Histogram Showing Shape of Raw Score Distribution Graphically

TITLE I - ARITHMETIC CONCEPTS - FALL 1969

Mean - 9.37

St.Dev. - 4.41

Raw Score	Cum. %	Stanine	Frequency	
28	99	9	1	*
27	99	9	2	**
26	99	9	0	
25	99	9	5	*****
24	98	9	5	*****
23	97	9	2	**
22	97	8	6	******
21	95	8	7	*******
20	94	8	11	***********
19	91	8	12	************
18	88	7	17	*****************
17	84	7	14	**************
16	81	7	15	***************
15	78	6	19	*******************
14	73	6	21	*********************
13	69	6	24	************************
12	63	5	34	**********************************
11	55	5	34	**********************************
10	48	5	43	***
9	38	4	36	************************************
8	30	4	40	**
7	20	3	35	***********************************
6	12	2	20	********************
5	8	2	14	**************
4	5	2	4	****
3	4	1	8	********
2	2	1	3	***
1	1	1	5	*****
			437	

FIGURE III-10
Frequency Distribution, Cumulative Percent Distribution, and Stanines Plus Histogram Showing Shape of Raw Score Distribution Graphically

TITLE I - ARITHMETIC CONCEPTS - SPRING 1970

Mean - 11.71

St.Dev. - 5.17

ANALYSES AND COMPARISON OF ITEM DIFFICULTY INFORMATION: TITLE I

The comprehensive item analysis information for Title I is given in Table III-2 in the R-W-O format. This table also includes the R/A ratio for convenience in comparing it with Rights or Wrongs.

Items having .00 in the "O" column are those which all children attempted. High percents of "R's" are associated with mastery of those items by most of the children.

This is a kind of general purpose table which is hard to interpret, especially in the Title I milieu.

The item analysis data for Word Meaning, Paragraph Meaning, and Arithmetic Computation, Concepts, and Applications is summarized in terms of Rights only in Table III-3 for the Title I population alongside similar data given for the random sample population, with which it is most appropriately compared.

It is from this comparative table, giving only the percent of items answered correctly, that we get the full impact of what these two groups of children knew at the beginning and what they had learned during the course of seven months of the school year.

Under normal circumstances in building a test of this sort, which is intended to be made up largely of material at grade or below grade in the curriculum sense (not difficulty), it is a common belief that one should be able to count on nearly half of the items, at least, having been mastered by the end of the period of instruction, or at the end of the fourth grade.

If we look in Table III-3 at the spring results for the random sample starting with item #16 under the Word Meaning column, we see one item in the spring passed by 74% of the pupils, but for the subsequent items the percents are far less than this.

Title I performance is disconcerting in that one doesn't find even 50% of the children answering any one question correctly at the end of the year until one goes back to item #13, and this leaves a very large proportion of the items which seem not to be measuring much.

We have to conclude that the movement of these percents toward mastery - i.e., 100% right - on this particular battery is woefully small. It is clear that this test simply was not an appropriate test to use for the Title I pupils at the fourth grade; even at the end of the fourth grade.

The item analysis shows the effect of instruction within the grade item by item, which is generally low, especially after the first twenty items. After allowing for an unearned bonus for guessing, much of the gains shown would, in fact, disappear.

Considering each item as an appropriate atom of information basic to progress in education and appropriately taught at the fourth grade or below, we must seek out the reasons for the small gains (in either group).

Was it lack of maintenance of skills instruction, or perhaps we should say "continued" instruction? Or was insufficient time spent on the subject in an already highly fractionated curriculum?

We can be sure there is no simple answer; but we also can be sure that the naive assumption that all were equally able to learn (if all were exposed to the same curriculum) was one basic reason.

In other words, the fault does not lie at the door of the fourth grade teacher alone or even mainly - it must be widely shared.

The fact that these children are in a Title I program is at least a recognition of this group as atypical and not properly measured against standards set for a universe of fourth graders.

Even allowing for the constantly reiterated differences between a test properly constructed for before and after testing compared with a general survey test, these results confirm the even greater unsuitability of Stanford: Intermediate I Battery: Form X for the job it was expected to do. Other data presented hereafter will confirm this conclusion.

Our most serious question now must be whether the same kind of result would have been obtained with some other survey battery similarly administered, intact, in the fall and spring as an evaluative device for measuring growth or learning, especially of individuals.

Table III-2

Percent Answering Each Item Right, Wrong, or Omit and Rights/Attempts for Each Item

TITLE I

ITEM	Fall	Word Meaning				Par. Meaning				Arith. Comp.				Arith. Concepts				Arith. Appl.			
No.	Spring	R	W	O	R/A	R	W	O	R/A	R	W	O	R/A	R	W	O	R/A	R	W	O	R/A
1	F	.55	.42	.03	.56	.71	.29	.01	.71	.81	.19	.00	.81	.54	.46	.00	.54	.65	.35	.00	.65
	S	.69	.30	.01	.70	.77	.23	.00	.77	.87	.13	.00	.87	.68	.32	.00	.68	.72	.28	.00	.72
2	F	.52	.46	.02	.53	.56	.44	.00	.56	.66	.33	.01	.67	.39	.58	.02	.40	.61	.38	.01	.62
	S	.69	.30	.01	.69	.67	.32	.00	.67	.70	.29	.01	.71	.45	.52	.03	.47	.64	.36	.00	.64
3	F	.40	.58	.02	.41	.64	.36	.00	.64	.68	.32	.00	.68	.46	.52	.02	.47	.45	.54	.01	.46
	S	.56	.42	.01	.57	.74	.26	.00	.74	.77	.22	.01	.77	.59	.41	.01	.59	.54	.45	.01	.55
4	F	.41	.56	.03	.42	.43	.56	.01	.43	.61	.38	.01	.61	.50	.48	.01	.51	.41	.58	.01	.42
	S	.59	.39	.02	.61	.52	.47	.00	.52	.72	.26	.02	.74	.57	.41	.02	.58	.38	.61	.00	.39
5	F	.46	.51	.02	.47	.44	.55	.01	.44	.56	.41	.03	.58	.35	.62	.02	.36	.35	.63	.02	.36
	S	.55	.43	.02	.56	.53	.46	.01	.53	.72	.27	.01	.73	.47	.50	.03	.48	.45	.54	.01	.46
6	F	.36	.61	.03	.37	.40	.58	.01	.41	.48	.49	.03	.50	.31	.65	.04	.33	.21	.77	.02	.22
	S	.50	.48	.02	.51	.54	.45	.01	.55	.76	.23	.01	.76	.34	.63	.03	.35	.36	.62	.01	.37
7	F	.48	.50	.02	.49	.37	.61	.02	.38	.53	.46	.01	.53	.30	.66	.04	.31	.46	.52	.02	.47
	S	.62	.37	.01	.63	.50	.49	.01	.51	.67	.32	.01	.68	.45	.51	.04	.47	.57	.42	.01	.57
8	F	.32	.64	.04	.33	.44	.53	.03	.45	.29	.70	.01	.29	.46	.51	.03	.47	.19	.78	.03	.20
	S	.50	.49	.01	.50	.56	.44	.00	.56	.46	.53	.01	.46	.58	.40	.01	.59	.28	.69	.03	.29
9	F	.31	.63	.06	.33	.43	.54	.03	.44	.36	.63	.01	.36	.25	.71	.04	.26	.42	.55	.03	.43
	S	.46	.52	.02	.47	.53	.46	.01	.54	.47	.51	.02	.47	.36	.61	.03	.37	.50	.48	.03	.51
10	F	.35	.59	.06	.37	.49	.48	.03	.50	.35	.62	.03	.36	.30	.66	.04	.32	.24	.72	.04	.25
	S	.53	.45	.02	.54	.57	.42	.01	.57	.50	.49	.01	.50	.36	.61	.03	.37	.30	.65	.05	.32
11	F	.28	.64	.08	.30	.45	.52	.03	.47	.17	.72	.11	.19	.35	.63	.02	.35	.23	.72	.05	.24
	S	.45	.53	.02	.46	.57	.42	.00	.57	.43	.51	.05	.46	.39	.60	.01	.39	.24	.73	.03	.25
12	F	.19	.73	.08	.20	.29	.68	.03	.30	.24	.68	.08	.26	.27	.70	.03	.28	.37	.58	.05	.39
	S	.33	.65	.02	.34	.37	.62	.01	.37	.50	.46	.04	.52	.33	.62	.04	.35	.39	.58	.02	.40
13	F	.38	.50	.12	.43	.50	.47	.03	.51	.27	.64	.09	.30	.18	.78	.04	.19	.23	.70	.06	.25
	S	.51	.45	.04	.53	.59	.40	.01	.59	.52	.44	.04	.54	.25	.73	.02	.25	.29	.69	.02	.30
14	F	.30	.59	.11	.33	.29	.67	.04	.30	.31	.61	.08	.34	.36	.60	.04	.37	.29	.65	.06	.31
	S	.38	.58	.04	.40	.37	.62	.01	.37	.38	.58	.04	.40	.44	.54	.02	.45	.32	.65	.03	.33
15	F	.24	.60	.16	.29	.31	.65	.04	.32	.31	.61	.08	.34	.32	.62	.06	.35	.21	.72	.07	.23
	S	.30	.64	.06	.32	.37	.62	.01	.37	.40	.55	.05	.42	.44	.53	.03	.45	.30	.67	.03	.31

Table III-2, Page 2- % Right, Wrong, or Omit and Rights/Attempts - Title I

ITEM No.	Fall Spring	Word Meaning				Par. Meaning				Arith. Comp.				Arith. Concepts				Arith. Appl.			
		R	W	O	R/A	R	W	O	R/A	R	W	O	R/A	R	W	O	R/A	R	W	O	R/A
16	F	.27	.56	.17	.33	.47	.49	.04	.48	.21	.71	.08	.23	.14	.79	.07	.15	.36	.58	.06	.38
	S	.41	.54	.05	.43	.57	.42	.01	.57	.36	.58	.05	.38	.21	.73	.06	.22	.50	.47	.03	.52
17	F	.26	.55	.19	.32	.23	.73	.04	.24	.24	.65	.11	.27	.30	.64	.06	.32	.31	.61	.08	.34
	S	.45	.48	.07	.48	.30	.69	.01	.30	.42	.52	.06	.45	.39	.57	.04	.41	.41	.54	.05	.43
18	F	.23	.56	.21	.29	.42	.52	.06	.45	.31	.48	.21	.39	.36	.56	.08	.39	.25	.67	.08	.27
	S	.38	.54	.08	.42	.53	.44	.03	.54	.44	.45	.11	.49	.45	.49	.06	.48	.34	.62	.04	.35
19	F	.18	.56	.26	.24	.28	.66	.06	.30	.29	.56	.15	.35	.33	.58	.09	.37	.24	.66	.09	.27
	S	.30	.58	.12	.34	.46	.53	.01	.46	.46	.45	.08	.51	.38	.57	.04	.40	.37	.58	.04	.39
20	F	.21	.51	.28	.30	.21	.73	.06	.22	.15	.66	.19	.18	.17	.74	.09	.19	.18	.71	.11	.20
	S	.33	.54	.13	.38	.29	.69	.02	.29	.24	.66	.10	.27	.22	.72	.06	.23	.30	.62	.08	.33
21	F	.26	.44	.30	.38	.33	.61	.06	.35	.09	.64	.27	.12	.28	.62	.10	.31	.28	.60	.12	.32
	S	.36	.51	.13	.41	.48	.50	.02	.49	.26	.61	.13	.30	.33	.61	.05	.35	.35	.59	.06	.37
22	F	.18	.51	.32	.26	.43	.51	.06	.46	.16	.54	.30	.22	.36	.55	.09	.40	.20	.67	.13	.23
	S	.30	.54	.16	.36	.53	.45	.02	.54	.24	.52	.23	.32	.43	.51	.06	.46	.24	.67	.08	.26
23	F	.17	.48	.35	.26	.36	.57	.06	.39	.12	.58	.30	.17	.27	.60	.13	.31	.15	.69	.16	.17
	S	.24	.56	.20	.30	.43	.55	.02	.44	.28	.56	.16	.33	.34	.57	.09	.38	.19	.71	.10	.21
24	F	.16	.46	.37	.26	.22	.70	.08	.23	.19	.51	.30	.27	.29	.56	.14	.34	.12	.72	.16	.14
	S	.27	.49	.24	.36	.25	.72	.03	.26	.28	.53	.19	.35	.31	.58	.10	.35	.13	.75	.12	.15
25	F	.19	.39	.42	.33	.20	.71	.09	.22	.08	.57	.35	.13	.28	.56	.15	.34	.21	.63	.16	.25
	S	.25	.49	.26	.34	.23	.73	.04	.24	.16	.56	.28	.22	.26	.64	.10	.29	.22	.67	.11	.25
26	F	.15	.41	.44	.27	.22	.29	.09	.24	.15	.48	.36	.24	.32	.50	.18	.40	.20	.62	.18	.24
	S	.31	.41	.28	.42	.24	.72	.04	.25	.25	.51	.24	.32	.36	.53	.11	.41	.26	.61	.13	.30
27	F	.13	.39	.48	.25	.23	.66	.11	.25	.12	.50	.38	.19	.08	.71	.21	.11	.11	.69	.20	.14
	S	.25	.43	.31	.37	.30	.65	.05	.31	.14	.55	.31	.20	.20	.67	.12	.23	.14	.72	.14	.17
28	F	.12	.38	.50	.23	.21	.66	.13	.24	.14	.48	.37	.23	.25	.51	.23	.33	.10	.68	.22	.13
	S	.18	.47	.35	.28	.31	.62	.07	.33	.30	.43	.27	.41	.37	.49	.14	.43	.12	.75	.13	.14
29	F	.08	.38	.54	.18	.23	.63	.13	.27	.13	.47	.39	.22	.09	.64	.27	.13	.09	.68	.23	.12
	S	.14	.46	.39	.23	.31	.62	.07	.33	.24	.46	.30	.34	.15	.69	.16	.18	.14	.71	.16	.16
30	F	.10	.33	.57	.24	.18	.67	.15	.21	.09	.49	.42	.15	.11	.64	.25	.15	.19	.58	.23	.25
	S	.23	.34	.43	.40	.19	.73	.08	.21	.18	.50	.32	.27	.15	.71	.14	.17	.21	.63	.16	.25

Table III-2, Page 3- % Right, Wrong, Omit and Rights/Attempts - Title I

ITEM No.	Fall Spring	Word Meaning				Par. Meaning				Arith. Comp.				Arith. Concepts				Arith. Appl.			
		R	W	O	R/A	R	W	O	R/A	R	W	O	R/A	R	W	O	R/A	R	W	O	R/A
31	F	.13	.28	.59	.31	.21	.62	.17	.25	.12	.46	.42	.21	.17	.57	.26	.23	.21	.52	.27	.29
	S	.14	.41	.45	.25	.29	.62	.09	.32	.19	.48	.33	.29	.21	.61	.18	.26	.18	.64	.18	.22
32	F	.11	.27	.62	.28	.34	.48	.18	.41	.08	.51	.41	.13	.15	.57	.28	.20	.14	.57	.29	.19
	S	.17	.34	.48	.34	.40	.50	.09	.45	.13	.56	.31	.18	.19	.63	.18	.24	.17	.64	.19	.21
33	F	.08	.29	.63	.21	.23	.59	.18	.28	.12	.44	.44	.22					.10	.61	.29	.14
	S	.16	.34	.50	.33	.30	.60	.10	.33	.16	.49	.35	.25					.15	.66	.19	.18
34	F	.06	.30	.64	.16	.31	.47	.22	.40	.10	.43	.47	.18								
	S	.15	.32	.53	.32	.37	.51	.12	.42	.09	.54	.37	.14								
35	F	.07	.27	.66	.21	.27	.50	.23	.35	.08	.48	.44	.14								
	S	.10	.35	.55	.22	.34	.53	.13	.39	.11	.52	.37	.18								
36	F	.08	.25	.67	.24	.28	.46	.25	.38	.05	.48	.47	.09								
	S	.09	.34	.57	.20	.35	.50	.15	.41	.07	.51	.42	.12								
37	F	.07	.24	.69	.23	.19	.53	.28	.27	.08	.44	.48	.15								
	S	.09	.32	.59	.22	.33	.52	.15	.39	.13	.46	.41	.22								
38	F	.07	.23	.70	.22	.20	.50	.30	.28	.10	.42	.48	.20								
	S	.07	.33	.60	.18	.31	.53	.16	.37	.12	.45	.42	.21								
39	F					.22	.46	.32	.32	.07	.44	.49	.13								
	S					.29	.54	.17	.35	.08	.50	.42	.14								
40	F					.21	.46	.33	.31												
	S					.26	.55	.19	.31												
41	F					.18	.48	.34	.28												
	S					.18	.62	.20	.22												
42	F					.24	.41	.35	.37												
	S					.35	.41	.24	.46												
43	F					.18	.43	.38	.29												
	S					.30	.45	.25	.40												
44	F					.18	.42	.40	.30												
	S					.28	.46	.26	.38												
45	F					.16	.43	.41	.27												
	S					.26	.48	.26	.35												
46	F					.12	.46	.42	.21												
	S					.20	.53	.27	.28												
47	F					.10	.47	.43	.18												
	S					.13	.58	.29	.18												
48	F					.17	.40	.43	.30												
	S					.22	.48	.30	.32												

ITEM No.	Fall Spring	Par. Meaning			
		R	W	O	R/A
49	F	.13	.40	.46	.25
	S	.14	.54	.32	.20
50	F	.07	.44	.49	.13
	S	.12	.54	.34	.18
51	F	.16	.34	.50	.32
	S	.20	.43	.36	.32
52	F	.07	.41	.52	.15
	S	.15	.47	.38	.23
53	F	.10	.38	.52	.21
	S	.20	.40	.39	.34
54	F	.07	.40	.53	.16
	S	.14	.45	.41	.24
55	F	.15	.31	.54	.33
	S	.20	.37	.45	.35
56	F	.15	.30	.54	.33
	S	.15	.39	.46	.28
57	F	.08	.37	.55	.19
	S	.13	.40	.47	.24
58	F	.12	.33	.55	.26
	S	.13	.39	.47	.25
59	F	.09	.35	.56	.21
	S	.11	.41	.48	.22
60	F	.12	.32	.56	.26
	S	.15	.37	.48	.28

Answer Sheet Study - III

Table III-3

A Comparison of the Percent Passing Each Item for the Random Sample Population and the Title I Population*

Item No.	Fall Spring	Word Meaning R.S.	Word Meaning Title I	Paragraph Meaning R.S.	Paragraph Meaning Title I	Arith. Comp. R.S.	Arith. Comp. Title I	Arith. Concepts R.S.	Arith. Concepts Title I	Arith. Appl. R.S.	Arith. Appl. Title I
1	F	.85	.55	.88	.71	.88	.81	.73	.53	.85	.65
	S	.93	.69	.93	.77	.88	.87	.85	.68	.86	.72
2	F	.77	.52	.80	.56	.77	.66	.51	.39	.83	.61
	S	.87	.69	.87	.67	.81	.70	.55	.45	.84	.64
3	F	.76	.40	.83	.64	.77	.68	.69	.46	.73	.45
	S	.88	.56	.93	.74	.85	.77	.75	.59	.76	.54
4	F	.72	.41	.66	.43	.79	.61	.61	.50	.62	.41
	S	.85	.59	.75	.52	.84	.73	.69	.57	.62	.38
5	F	.71	.46	.65	.44	.74	.56	.46	.35	.59	.35
	S	.82	.55	.77	.53	.87	.72	.61	.47	.68	.45
6	F	.60	.36	.68	.40	.61	.48	.33	.31	.36	.21
	S	.72	.50	.81	.54	.85	.76	.38	.34	.51	.36
7	F	.78	.48	.61	.37	.67	.53	.41	.30	.68	.46
	S	.87	.62	.78	.50	.77	.67	.63	.45	.75	.57
8	F	.66	.32	.66	.44	.39	.29	.61	.46	.29	.19
	S	.82	.50	.71	.56	.63	.46	.67	.58	.46	.28
9	F	.65	.31	.68	.43	.46	.36	.31	.25	.61	.42
	S	.79	.46	.80	.53	.65	.47	.50	.36	.66	.50
10	F	.60	.35	.75	.49	.45	.35	.43	.30	.29	.24
	S	.75	.53	.81	.57	.65	.50	.45	.36	.48	.30
11	F	.49	.28	.77	.45	.23	.17	.53	.35	.31	.23
	S	.69	.45	.83	.57	.58	.43	.53	.39	.45	.24
12	F	.43	.19	.41	.29	.29	.24	.48	.27	.61	.37
	S	.59	.33	.57	.37	.65	.50	.59	.33	.67	.39
13	F	.73	.38	.66	.50	.41	.27	.17	.18	.41	.23
	S	.83	.51	.75	.59	.71	.52	.33	.25	.53	.29
14	F	.57	.30	.55	.29	.35	.31	.57	.36	.49	.29
	S	.65	.38	.70	.37	.53	.38	.67	.44	.57	.32
15	F	.43	.24	.38	.31	.35	.31	.53	.32	.41	.21
	S	.65	.30	.46	.37	.48	.40	.63	.44	.47	.30
16	F	.63	.27	.71	.47	.27	.21	.22	.14	.59	.36
	S	.74	.41	.79	.57	.51	.36	.34	.21	.70	.50
17	F	.59	.26	.31	.23	.39	.24	.52	.30	.51	.31
	S	.69	.45	.42	.30	.54	.42	.64	.39	.59	.41
18	F	.41	.23	.63	.42	.32	.31	.37	.36	.47	.25
	S	.57	.38	.76	.53	.65	.44	.61	.45	.62	.34

*Random Sample: N - 567   Title I: N - 426

Table III-3 - page 2

Item No.	Fall Spring	Word Meaning R.S.	Word Meaning Title I	Paragraph Meaning R.S.	Paragraph Meaning Title I	Arith. Comp. R.S.	Arith. Comp. Title I	Arith. Concepts R.S.	Arith. Concepts Title I	Arith. Appl. R.S.	Arith. Appl. Title I
19	F	.39	.18	.57	.28	.37	.29	.55	.33	.40	.24
	S	.61	.30	.71	.45	.59	.46	.61	.38	.55	.37
20	F	.45	.21	.43	.21	.25	.15	.17	.17	.25	.18
	S	.59	.33	.58	.29	.43	.24	.30	.22	.52	.30
21	F	.43	.26	.66	.33	.08	.09	.43	.28	.46	.28
	S	.55	.36	.77	.48	.48	.26	.53	.33	.59	.35
22	F	.33	.18	.66	.43	.17	.16	.60	.36	.33	.20
	S	.47	.30	.77	.53	.33	.24	.69	.43	.43	.24
23	F	.32	.17	.53	.36	.15	.12	.45	.27	.23	.15
	S	.49	.24	.61	.43	.46	.28	.57	.34	.39	.19
24	F	.34	.16	.29	.22	.17	.19	.33	.29	.11	.12
	S	.60	.27	.39	.25	.38	.28	.41	.31	.26	.13
25	F	.35	.19	.26	.20	.07	.08	.41	.28	.24	.21
	S	.62	.25	.28	.23	.23	.16	.49	.26	.23	.22
26	F	.37	.15	.19	.22	.13	.15	.43	.32	.21	.20
	S	.61	.31	.28	.24	.34	.25	.63	.36	.33	.26
27	F	.21	.13	.43	.23	.09	.12	.13	.08	.17	.11
	S	.40	.25	.41	.30	.18	.14	.23	.20	.27	.14
28	F	.19	.12	.38	.21	.13	.14	.43	.25	.04	.10
	S	.39	.18	.52	.31	.43	.30	.59	.37	.17	.12
29	F	.09	.08	.42	.23	.09	.13	.13	.09	.09	.09
	S	.30	.14	.54	.31	.29	.24	.22	.15	.21	.14
30	F	.21	.10	.31	.18	.07	.09	.05	.11	.24	.19
	S	.37	.23	.43	.19	.24	.18	.17	.15	.33	.21
31	F	.17	.13	.41	.21	.09	.12	.13	.17	.20	.21
	S	.38	.14	.51	.29	.27	.19	.21	.21	.33	.18
32	F	.13	.11	.47	.34	.05	.08	.19	.15	.09	.14
	S	.29	.17	.62	.40	.25	.13	.23	.19	.15	.17
33	F	.15	.08	.35	.23	.07	.12			.09	.10
	S	.38	.16	.45	.30	.22	.16			.16	.15
34	F	.13	.06	.45	.31	.05	.10				
	S	.27	.15	.64	.37	.07	.09				
35	F	.08	.07	.37	.27	.06	.08				
	S	.25	.10	.53	.34	.16	.11				

Answer Sheet Study - III

Table III-3 - page 3

Item	Fall	Word Meaning		Paragraph Meaning		Arith. Comp.		Arith. Concepts		Arith. Appl.	
No.	Spring	R.S.	Title I	R.S.	Title I	R.S.	Title I	R.S.	Title I	R.S.	Title I
36	F	.05	.08	.45	.28	.03	.05				
	S	.16	.09	.61	.35	.10	.07				
37	F	.06	.07	.42	.19	.05	.08				
	S	.22	.09	.65	.33	.13	.13				
38	F	.05	.07	.41	.20	.05	.10				
	S	.13	.07	.62	.31	.11	.12				
39	F			.31	.22	.09	.07				
	S			.52	.29	.13	.08				
40	F			.39	.21						
	S			.57	.26						
41	F			.21	.18						
	S			.39	.18						
42	F			.37	.24						
	S			.59	.35						
43	F			.35	.18						
	S			.59	.30						
44	F			.25	.18						
	S			.43	.28						
45	F			.27	.16						
	S			.47	.26						
46	F			.17	.12						
	S			.35	.20						
47	F			.10	.10						
	S			.23	.13						
48	F			.27	.17						
	S			.47	.22						
49	F			.13	.13						
	S			.23	.14						
50	F			.05	.07						
	S			.10	.12						
51	F			.18	.16						
	S			.35	.20						

Item	Fall	Paragraph Meaning	
No.	Spring	R.S.	Title I
52	F	.12	.07
	S	.33	.15
53	F	.19	.10
	S	.42	.20
54	F	.11	.07
	S	.26	.14
55	F	.15	.15
	S	.31	.20
56	F	.16	.15
	S	.35	.15
57	F	.06	.08
	S	.15	.13
58	F	.09	.11
	S	.23	.13
59	F	.07	.09
	S	.11	.11
60	F	.09	.11
	S	.17	.15

ANALYZING THE FALL-SPRING PUPIL RESULTS BY CATEGORY (BY WAY OF REVIEW)

The fact that the same test was administered twice, once in the fall and once in the spring, gave us an extraordinary opportunity to compare results item by item. We can now see how a child answers a particular item in the fall and also in the spring. This has proved to be a very enlightening investigation.

The analysis has been approached from two different points of view. In both cases we start with what might be called the pupil-item response. For example, in Word Meaning there are a total of 38 items and 426 Title I pupils, boys and girls combined. 38 x 426 gives us the surprising total of 16,188 sets of pupil-item responses which could have been marked by these particular pupils.

A study of these raw data led us to set up comprehensive categories into which they might be divided. An example is a child who marked an item Right in the fall and Right in the spring. This is the "RR" category in our Table III-4. Another example is a child who marked any response Right in the fall and marked it Wrong in the spring ("RW").

It is essential to remember that any pupil in the population marking any item Right in the fall and Right in the spring ("RR") falls in that category for that item only. This does not mean that he habitually marks items Right in the fall and Right in the spring. It might conceivably be the only item he answered in this fashion.

The same would be true, of course, of the other categories. All of the "RR" responses, regardless of who made them or how many were made by whom, is first expressed as a percent of all of the items marked Right/Right, etc. (Table III-4)

All "RR" items obviously serve no measurement function in evaluating gains, but do establish an evaluation base for specific items; i.e., these are items not requiring additional instruction.

Another approach to the interpretation of these data, as you may recall from Part II, was to express the results of each category as a mean, dividing the total number of pupils marking "RR" by the number of pupils in the population, and so on for all categories. (Table III-5)

It would be foolish to spend time making detailed comparisons from one population to the other in such a complex situation. The data by category is made available in this report for detailed comparative study for any who wish to make the effort.

Let us say by way of summary that the tables for Title I yielded no surprises as respect to the comparative performance of Title I and the random sample.

The two population tables are not the same by any means, as one can see by looking at the mean for Right/Right responses, for example, first for the random sample and then for Title I. (Tables II-3 and III-5)

The average number of items in the Word Meaning "RR" category for the random sample was 13 compared to 5 for Title I. 5.6 is the mean of the "WR" responses for the random sample; for Title I, it is 5.9, a very interesting difference suggesting a rather more effective instruction for Title I than the random sample (or better guessing)!

Table III-6 provides an interesting summary. The "WR + OR" section is a kind of measure of the positive effects of all learning experiences over seven months of school; the "RW + RO" is a kind of complement or reversal, suggesting either forgetting or the operation of chance on the first test.

In Table III-6 the "happy" side is clearly shown in the "WR + OR" means; e.g., an average change to "R" of 8.1 points in Word Meaning and 8.4 points in Computation.

The less salubrious finding of average changes from "R" to "W" or "O" reveal the instability of item performance with the obvious choices of either forgetting or guessing as the villians.

Comparisons with the random sample here are in favor of Title I, although absolute values are mostly higher in the representative group. Title I gets a plus.

Perhaps we set our hopes too high as to how much learning takes place in a specific subject matter field when we recall the obvious fact that the child's school day amounts to only from four to six or seven hours, depending on grade level, and that these hours are broken up into many other things other than instruction in specific school subjects; such as physical education, recess periods, arts and crafts, directed play activities, pageants, etc.

Most schools also allow a certain amount of time for a luncheon period, so the actual amount of time left for instruction does not allow a very large amount of time per day per subject.

Table III-4

Number and Percent of Pupil-Item Responses by Consistency Category

TITLE I GROUP of 426 Boys and Girls

Test	No. of Items	Possible Fall-Spring Responses	MARKED RESPONSES*					O M I T S**				
			RR	RW	WR	WWS	WWD	OO	OR	OW	RO	WO
Word Meaning	38	16,188	2110	1387	2515	1594	2491	2133	939	1828	292	899
Percent of Responses			13.0	8.6	15.5	9.8	15.4	13.2	5.8	11.3	1.8	5.6
Paragraph Meaning	60	25,560	3407	3062	4238	2839	4847	2473	1135	2152	380	1027
Percent of Responses			13.3	12.0	16.6	11.1	19.0	9.7	4.4	8.4	1.5	4.0
Arithmetic Computation	39	16,614	2469	1509	2872	1475	3204	1515	712	1569	236	1053
Percent of Responses			14.9	9.1	17.3	8.9	19.3	9.1	4.3	9.4	1.4	6.3
Arithmetic Concepts	32	13,632	2069	1766	2599	2582	2761	358	307	672	135	383
Percent of Responses			15.2	13.0	19.1	18.9	20.3	2.6	2.2	4.9	1.0	2.8
Arithmetic Applications	33	14,058	1820	1803	2461	2002	3944	471	292	758	112	395
Percent of Responses			12.9	12.8	17.5	14.2	28.1	3.4	2.1	5.4	.8	2.8

*RR-Right Fall and Spring
RW-Right Fall-Wrong Spring
WR-Wrong Fall-Right Spring
WWS-Wrong Fall and Spring-Same Response
WWD-Wrong Fall and Spring-Different Response

**OO-Omit Fall and Spring
OR-Omit Fall-Right Spring
OW-Omit Fall-Wrong Spring
RO-Right Fall-Omit Spring
WO-Wrong Fall-Omit Spring

Table III-5

Number and Mean of Pupil-Item Responses by Consistency Category

TITLE I GROUP of 426 Boys and Girls

Test	No. of Items	Possible Fall-Spring Responses	MARKED RESPONSES*					O M I T S**				
			RR	RW	WR	WWS	WWD	OO	OR	OW	RO	WO
Word Meaning	38	16,188	2110	1387	2515	1594	2491	2133	939	1828	292	899
Mean Responses			5.0	3.3	5.9	3.7	5.8	5.0	2.2	4.3	.7	2.1
Paragraph Meaning	60	25,560	3407	3062	4238	2839	4847	2473	1135	2152	380	1027
Mean Responses			8.0	7.2	9.9	6.7	11.4	5.8	2.7	5.0	.9	2.4
Arithmetic Computation	39	16,614	2469	1509	2872	1475	3204	1515	712	1569	236	1053
Mean Responses			5.8	3.5	6.7	3.5	7.5	3.6	1.7	3.7	.5	2.5
Arithmetic Concepts	32	13,632	2069	1766	2599	2582	2761	358	307	672	135	383
Mean Responses			4.9	4.1	6.1	6.1	6.5	.8	.7	1.6	.3	.9
Arithmetic Applications	33	14,058	1820	1803	2461	2002	3944	471	292	758	112	395
Mean Responses			4.3	4.2	5.8	4.7	9.2	1.1	.7	1.8	.3	.9

*RR-Right Fall and Spring
RW-Right Fall-Wrong Spring
WR-Wrong Fall-Right Spring
WWS-Wrong Fall and Spring-Same Response
WWD-Wrong Fall and Spring-Different Response

**OO-Omit Fall and Spring
OR-Omit Fall-Right Spring
OW- Omit Fall-Wrong Spring
RO- Right Fall-Omit Spring
WO- Wrong Fall-Omit Spring

Table III-6

Analysis of Categories "WR+OR"* and "RW+RO"**

TITLE I

Test	No. of Items	Possible Pupil-Item Responses	Selected Pupil-Item Responses "WR+OR" No.	Mean	%	"RW+RO" No.	Mean	%
Word Meaning	38	16,188	3454	8.1	21	1679	3.9	10
Paragraph Meaning	60	25,560	5373	12.6	21	3442	8.1	13
Arithmetic Computation	39	16,614	3584	8.4	22	1745	4.1	11
Arithmetic Concepts	32	13,632	2906	6.8	21	1901	4.5	14
Arithmetic Applications	33	14,058	2753	6.5	20	1915	4.5	14

* Wrong or Omit fall, Right spring = possible gain

** Right fall, Wrong or Omit spring = possible fall guessing

When one further allows for holiday forgetting, especially during the long summer vacation (thus requiring a major maintenance of skills development in the fall in order that the children may continue to know what they have "known" at one time), there is not very much time left for positive instruction in new content.

In summary, note that only Wrong/Right items or Omit/Right items can result in a gain in score. This suggests that the fall test be limited only to "hard" items, with plenty of room for more of these pleasant positive changes over the school year - IF CONDITIONS FOR LEARNING ARE FAVORABLE.

Even when one examines such data as these for those items which are specifically restricted to the content previously accepted by the teachers as their responsibility during the school year, the gains are not proportionally higher (in this writer's experience).

It still takes a period of at least three years for the average child to reach a point of mastery 75% of the time for a family of interchangeable items, such as the addition of two three-place numbers.

We must see how unrealistic it is for the criterion reference advocates to purport to see an hierarchy in every subject and to require that every child shall show substantial mastery at any one level before he's permitted to go on to the next.

This is especially true when the majority of schools in this country still fail to recognize, overtly at least, the tremendous range of talent to be found within any grade.

If this statewide testing program accomplished no other purpose, it should have demonstrated to those who examined all of the results carefully, either at the local or at the state level, that the range of talent to be found in the average grade (grade four, in this case) is simply enormous.

Considering this wide range in pupil ability and the astonishing variability in the training and competence of a teacher to teach all of the subjects to which she may be assigned in the elementary grades, the results look less dismal for the tests but more dismal for the education process.

The almost total absence of pre-service instruction in the training of teachers in the evaluative processes, techniques, and procedures - including the construction of good and valid tests of their own - makes it all the more surprising that children get "sorted out" as well as they do. Education is, perhaps far more than we are willing to concede, a pupil-managed experience.

We must press on to other approaches to these data and to continue to search diligently for usable and actionable kinds of TRUTH concerning individual children which will help the teachers and all those who deal with them (including parents and administration) to make the learning experience more effective, both collectively and individually.

Summary

In this section we have clearly established the existence of high levels of guessing on multiple choice material. This certainly suggests that pupils take the attitude that tests are punitive: low score = failure; high score = praise and a reputation for excellence.

The extremely low correlations between fall and spring results in both populations show the effects of this behavior and nullify much of the good that tests can surely do in improving instruction.1/

1/ See Table VII-B-1 of the first Title I Report for Fall vs Spring correlations for Random Sample and Title I.

RIGHTS VERSUS ATTEMPTS FOR TITLE I AS COMPARED TO THE RANDOM SAMPLE

In discussing the data for Title I, it is more than ever important to consider the relationship between the number of items attempted and the number attempted which are answered correctly.

In Part II of this report, in which the statistical data concerning the random sample are discussed, we have attacked this problem in a number of ways. We started with the correlation between Rights and Attempts, which we assumed would be high but wasn't.

In Table II-7 of Part II we have given correlations together with the means and standard deviations for the Attempts and Rights for the total random sample.

In Table III-7 of this section, we give data for the 426 case Title I sample, boys and girls combined (sex differences not having seemed to be important in this instance).

The Title I correlations are almost always lower. There are one or two exceptions, one being the .36 reported for Arithmetic Concepts in the fall as compared to .26 for the random sample.

There is another exception for the same test administered in the spring, where the correlation is .37 as compared to .33 in the random sample. These differences could be chance alone and certainly are not significant.

The truth of the matter is that the relationship between Rights and Attempts is woefully low in terms of what one would expect in tests of the sort we have been studying. This is especially so when the test was intended by Title I to be a serious evaluation of what children know at the beginning in these basic fields covered by this particular study and what they learn over a seven-month period.

Correlations between Rights and Attempts for the Title I sample of the magnitude shown immediately raise doubts as to the validity of the whole effort.

(This generalization includes the somewhat higher but still alarmingly low correlation for the random sample. It is positive proof that guessing is tending to invalidate the data.)

The Coefficient of Alienation

Caution: Correlation coefficients are not percentages; absence of relationship is not measured by considering the complement of the correlation.

The measure of lack of relationship corresponding to the correlation coefficient is the coefficient of alienation.

In this regard, a word of wisdom from Psychometric Methods, by J. P. Guilford, 1936 edition, page 362, is especially clear. (Some minor modifications needed for clarity made by the writer as indicated.)1/

"If r (r being the symbol for correlation) is equal to 1.00, then the expression $\sqrt{1-r^2}$ *(k) is equal to zero and the error of prediction of y from x is nil. If on the other hand r is zero, then the expression "k" is equal to 1.00. In this case, the errors of prediction of Y are as great as the errors could be even without any knowledge of X to help in predicting Y.*

"The coefficient of alienation (k) measures the absence of relationship just as r (or the correlation coefficient) measures its presence."

(For a correlation value of .43, the corresponding coefficient of alienation is .903. The percentage of reduction in the errors of prediction of y from x, which is of major concern to the test user, amounts under these circumstances to only 9.7.)

Is the use of this quotation intended to be an indictment of testing or even an indictment of the Stanford Achievement Test? Certainly not. It may be that the training of teachers is remiss in failing to highlight the dangers of guessing.

The correlations of Rights versus Attempts are so low (considering the information given by the coefficient of alienation) that one can say that, for all practical purposes, it is impossible to predict with a significant degree of precision how many items a child would attempt if one knew only how many he had answered correctly.

Woe is the condition of education when a child depends upon random guessing to earn a score that he hopes will satisfy his teacher and the school administration; or perhaps prevent his being labeled as a failure. This fear of failure, indeed, is very often the best motivation for testing in the fall before the pupil's feeling of needing to succeed becomes reborn and strengthened by his school experience between the beginning and close of school.

1/ Guilford, J. P. Psychometric Methods. New York: McGraw-Hill, 1936, 362.

Table III-7

Attempts versus Rights
Correlations, Means, and Standard Deviations

TITLE I

Test	r	k*	Mean Attempts	Mean Rights	Standard Dev. Attempts	Standard Dev. Rights
FALL:						
Word Meaning	.50	.87	25.55	8.72	9.41	4.86
Paragraph Meaning	.43	.90	46.67	16.23	14.96	5.84
Arithmetic Computation	.27	.96	29.91	9.89	10.08	4.07
Arithmetic Concepts	.36	.93	28.86	9.32	5.74	4.30
Arithmetic Applications	.25	.97	29.35	8.77	6.71	4.16
SPRING:						
Word Meaning	.44	.90	29.93	13.06	8.47	6.16
Paragraph Meaning	.26	.97	51.13	20.67	12.42	7.83
Arithmetic Computation	.35	.94	32.05	14.21	9.11	6.45
Arithmetic Concepts	.37	.93	29.72	11.68	5.43	5.18
Arithmetic Applications	.26	.97	30.32	10.73	6.25	5.59

* k=Coefficient of Alienation

Distribution of Attempts

Let us now consider the distribution of attempts, as was done previously for the random sample.1/

Notice first in Table III-8 the large, even dismaying, proportion of those attempting all items. This is the number which appears at the top of each column and is boxed to make it stand out more prominently.

(The minor differences in the number of cases included in the several distributions from subject to subject are due only to the fact that all tests were not administered on the same day, and a few children were absent on at least one of the testing days in the fall or spring.)

Note: We are not dealing with number right scores here, but merely attempts. These numbers are further highlighted at the bottom of the table, where the percent of the total number of pupils tested who marked all items is given.

For example, 25% of children in the Title I group taking the Word Meaning Test in the fall marked all 38 response spaces, even though the distribution of scores (number right) shows that this test is on the hard side for this population; very few pupils earn scores anywhere near a perfect score.

In the case of the Title I population, the proportion of children attempting all items goes up slightly from fall to spring, but only slightly; contrariwise, the children in the random sample show a substantially larger percentage marking all items in the spring, especially in the Word Meaning and Paragraph Meaning Tests.

The trend is clearly noticeable for the other tests, but to a lesser degree, and is a normal expectancy since the random sample group starts at a higher level - leaving fewer items on which there is a need to guess.

Perhaps we should stop at this point to consider again why we are concerned about the large numbers in the Attempted-All category. Does it not indicate that there is a terrible lack of mutual confidence on the part of the pupil and the teacher?

What happens to the ethical considerations involved? In using these test results, when the child feels he must resort to random marking of all or many items in the hope of picking up a few additional points by chance alone, is the whole test invalidated?

Does the pupil want to fool the teacher into thinking he has accomplished more than he has in fact accomplished? If so, why?

Is the equating of a low score with failure due to willful neglect on the part of the teacher to consider fully his potential?

Guessing may be a natural tendency encouraged by the item type, but one cannot help but feel that it is indeed tied up inextricably with the real or implied punitive use of the test scores.

This idea is abhorrent to any professional test maker or expert in the field of the use of tests as a function of instruction.

Analysis of the Scores of the "Attempted-All" Group

At this point it becomes logical to analyze the actual scores of those individuals who attempted all items as compared to the distribution of scores for the total sample. This was done for the random sample and discussed in detail.2/

Reproduced in Table III-9 are similar distributions of scores for those children in Title I who Attempted-All items. The numbers of children in this category are given at the bottom of the table. Percents are also given and justify careful study.

All of these tests are short, being in the order of 32 to 39 items with the exception of Paragraph Meaning. This may have encouraged larger numbers of random answers, although the distributions of scores for all pupils (as you will recall from our earlier discussion) show that the tests are on the difficult side for Title I cases, especially the Concepts and Applications Tests - perhaps because these tests involve a large degree of logical reasoning compared to rote learning.

It is noteworthy in studying this table to see that the scores for the Attempted-All group cover most of the range of scores possible in the test.

In discussing this phenomenon in connection with the random sample in Part II, we pointed out that in some cases Attempted-All did not mean guessing; it simply meant that some were really able to answer all or nearly all of the items.

1/ See Table II-8, Part II.

2/ See Part II, Table II-9.

Table III-8

Distribution of Attempts with Means and Standard Deviations

TITLE I Group of 426 Boys and Girls

No. of ATTEMPTS	Word Meaning		Paragraph Meaning		Arithmetic Computation		Arithmetic Concepts		Arithmetic Applications	
	F	S	F	S	F	S	F	S	F	S
60			145	184						
59			17	26						
58			18	7						
57			3	5						
56			6	7						
55			7	6						
54			1	11						
53			3	10						
52			4	5						
51			4	8						
50			6	7						
49			8	10						
48			13	8						
47			5	5						
46			2	8						
45			4	2						
44			9	3						
43			7	2						
42			6	6						
41			11	15						
40			5	7						
39			6	6	155	181				
38	107	145	8	4	34	31				
37	13	18	9	2	8	15				
36	12	12	4	4	8	8				
35	5	10	12	9	12	12				
34	10	9	7	1	4	8				
33	5	9	16	8	8	12			261	301
32	7	12	4	6	10	8	259	303	24	33
31	8	3	6	4	6	6	32	20	12	4
30	11	10	1	3	3	4	8	21	16	8
29	14	12	8	7	4	11	7	6	7	5
28	19	16	6	2	12	7	16	6	11	6
27	9	14	8	5	5	7	13	11	7	5
26	13	15	6	5	7	8	10	6	7	4
25	8	14	3	1	10	13	14	3	7	4
24	18	13	2		6	7	4	5	10	4
23	11	13	5	4	8	16	8	7	5	7
22	14	15			10	7	8		6	8
21	8	12	1	3	13	6	3	2	2	3
20	14	9	4	2	9	13	7	8	6	5
19	11	11	1	1	18	5	2	4	3	5
18	15	10	1		10	5	8	2	4	5
17	11	5	2	1	16	5	3	3	2	3
16	5	5	1		10	4	3	4	7	1
15	15	4	3	1	10	1	5	1	1	1
14	9	3			4	3	2	3	4	1
13	9	4	3		1	3			6	1
12	8	3	1		4	3	4		3	
11	7	2	3		5	4	3	1	2	
1-10	28	5	9	2	14	9	7	7	12	7
N	424	423	424	423	424	422	426	423	425	421
Mean	26.7	30.4	46.5	51.1	30.3	32.6	29.2	30.2	29.7	31.0
St.Dev.	10.3	8.5	15.1	12.4	10.0	9.1	5.8	5.1	6.7	5.4
% Att.All	25%	34%	34%	43%	37%	43%	61%	72%	61%	71%

Table III-9

Distribution of Right Responses for Students Who Attempted All Items

TITLE I Group of 426 Boys and Girls

No. of RIGHTS	Word Meaning 38 ATTS.		Paragraph Meaning 60 ATTS.		Arithmetic Computation 39 ATTEMPTS		Arithmetic Concepts 32 ATTS.		Arithmetic Applications 33 ATTEMPTS	
	F	S	F	S	F	S	F	S	F	S
54				1						
53										
52										
51										
50										
49										
48										
47										
46										
45										
44				4						
43										
42				1						
41				1						
40				1						
39										
38				1						
37						1				
36				1		1				
35				3						
34		1		6		1				
33		1	1	1						
32	1	1		3		1				1
31		1	1	2		1				
30	1			4		3				
29		4	1	3		2				1
28			1	3		2		1		
27		2	1	3		2	1	2		
26		3	3	4		2				3
25		2	4	5		2	1	2		1
24		3	3	12		5		4	1	5
23	1	3	3	8		4	1	2		3
22	1	7	9	14		4		5		6
21		3	9	9	2	2		6		3
20	1	4	6	11		9	4	9	2	9
19	3	2	8	7	4	5	4	10	2	3
18	4	9	19	11	5	6	8	14	4	7
17	1	6	11	7	5	4	3	12	6	9
16	1	8	16	13	3	9	6	7	9	11
15	2	8	12	7	8	13	11	14	7	10
14	5	7	12	9	5	13	11	16	12	13
13	5	12	10	6	10	13	14	15	12	23
12	13	7	8	10	11	10	17	27	18	13
11	7	3		2	11	16	24	27	10	18
10	15	17	2	4	13	17	25	31	19	20
9	11	15	2	3	20	5	28	28	28	26
8	15	7	1	1	17	7	20	23	26	25
7	9	3		2	14	11	30	23	31	33
6	4	4	2		17	6	23	11	26	13
5	4	1		1	5	2	18	9	22	19
0-4	3	1			5	2	10	5	26	26
Total N	107	145	145	184	155	181	259	303	261	301
%ofTi.I	25%	34%	34%	43%	37%	43%	61%	72%	61%	71%
Mean	11.0	15.3	17.6	21.4	10.5	15.0	10.1	12.3	9.2	11.1
St.Dev.	4.9	6.4	4.7	8.5	4.1	6.8	4.2	5.1	4.0	5.6

In such cases, the gap between the earned score and the number of attempts was small, as it should be, even though all items were marked.

It is quite obvious that if one associates previous low achievement with appropriate assignment to a Title I project, many of these children do not belong in Title I. This was even more clearly seen when we were discussing the distributions of scores in histogram form with stanines and cumulative percentile ranks given.

The question now is, of course, what is the significance of all this?

We have demonstrated, first, that guessing does exist; secondly, that even though the amount of such guessing by child A or B versus W, Y, or Z may not affect the group totals greatly, it may be disastrous for the individual if he is allowed to receive credit where credit is not due.

We have also seen that this is especially disastrous when one considers the item analysis data and realizes that one can no longer accept a right response as truly indicating knowledge.

We have seen that our survey instrument was not appropriate; if Stanford can be considered to be typical, and the behavior of New Hampshire children characteristic of children everywhere, as it certainly should be, this may be considered a generalization.

The type of before-after testing done here has been greatly encouraged by our federal programs, such as the emphasis on accountability in the OEO programs and in USOE, especially Title I.

We have had to come to the conclusion, especially after examining the individual protocols, that children are so different in their test-taking performance that no uniform correction for guessing is possible.

How pupils behave depends upon the length of the test, the nature of the subject matter, the level of mastery at the beginning of previously taught material, the urge or drive to do well in the sight of his authority figures (or absence of that urge), and many other reasons not determined.

Each child will perform in a manner which is his individual reaction to the total situation, determined not only by intellect or knowledge but also by his emotional condition and his temperament as well.

If any generally applied formula for correction for guessing is bound to fail, what then is left for us?

Basically, the answer must be a total reexamination of the evaluation procedure - over the short range certainly, and only to a lesser degree over the long range.

Subject matter does not fall into neat categories or hierarchies; most such hierarchies are constructs only existing in the minds of the textbook writer and/or the experts who may produce a yearbook or some other standard work which is supposed to establish for everyone what the pattern of learning is to be, both as regards the content and the order in which this content is presented.

THE R/A RATIO AS APPLIED TO TITLE I

In Part II of this report, the development of the R/A ratio is described in detail; hence there is no need for repetition here, especially since the concept is so elementary. The theory that R/A should be high is based upon the belief that guessing or random marking has no place in a valid testing procedure.

The Title I R/A data for each item are given in the same table that gives R-W-O data (Table III-2). Further information is provided by Chart III-1, a graph on normal probability paper of the actual distribution of R/A for Word Meaning, Title I.

This doesn't look much different than the similar chart for the random sample. The range is comparable - more or less an artifact by the nature of the statistic. The median R/A is lower; i.e., it moves away from 1.00 more than was true for the random sample (.60 versus .31 for the fall testing).

In Part II, we moved away from the position that the R/A ratio was an effective statistic in all cases because of the limitations placed upon it by the incidence of very high or very low scores, which restrict or modify the range of R/A.

The recommendation was made that this ratio should be considered as another bit of confirmatory evidence when substantial guessing was indicated by a large number of wrong responses. It is most dependable for the scores falling within the range of the mean plus or minus one standard deviation, or for stanines in the 4-5-6 group - a somewhat more restrictive application.

The R/A ratio is roughly comparable from test to test, whereas raw scores are not. Thus, it has the effect of helping us determine on a universal basis how small a ratio can be considered acceptable.

It is a statistic that belongs primarily in a compilation of data for one child: including Rights, Wrongs, Omits; the incidence of RR's, WR's, etc., in the categorical analysis; etc.

The biggest problem is to arrive at some conclusion, not statistically but logically, as to how much the ratio can depart from 1.00 and not cause alarm.

In a situation where random responses seem of almost epidemic proportion, certainly the average R/A is no infallible guide.

On the other hand, the establishment of 1.00 as the norm ignores the "dare" to take a chance made by the multiple choice-type question, especially in a situation where one or two options can be logically eliminated.

In determining the acceptable lower limit (the higher the better), one must remember that random marking reduces the correlation between before-after tests in strict proportion to the use of guessing as a response technique.

This miasma of half-knowledge, inevitable in a situation where accepted truth one day is often reversed the next, makes acting on the basis of "best evidence" a necessity of life. A logical deduction cannot be rejected; neither can it be identified from the purely random response.

There is no absolute answer to this dilemma short of adding to each response set an estimate by the pupil of the degree of "sureness" with which he has answered each question. Even in this situation, temperamental factors influence each pupil's response differentially.

Raw Score	Cum. %	Stanine	Frequency	
31	99	9	1	*
30	99	9	0	
29	99	9	0	
28	99	9	0	
27	99	9	1	*
26	99	9	1	*
25	99	9	2	**
24	99	9	4	****
23	98	9	11	***********
22	97	8	8	********
21	95	8	13	*************
20	93	8	23	***********************
19	89	7	25	*************************
18	85	7	22	**********************
17	81	6	44	**
16	73	6	29	*****************************
15	68	6	39	***************************************
14	62	5	37	*************************************
13	55	5	29	*****************************
12	51	5	41	*******************************
11	43	4	46	**
10	36	4	43	***
9	28	4	49	***
8	20	3	30	******************************
7	15	3	31	*******************************
6	9	2	24	************************
5	5	2	14	**************
4	3	1	8	********
3	2	1	5	*****
2	1	1	4	****
			584	

Appendix A5
Frequency Distribution, Cumulative Percent Distribution, and Stanines
Plus Histogram Showing Shape of Raw Score Distribution Graphically

RANDOM SAMPLE - ARITHMETIC APPLICATIONS - FALL 1969
Mean - 12.93
St.Dev. - 5.03

Answer Sheet Study - III

Chart III-1

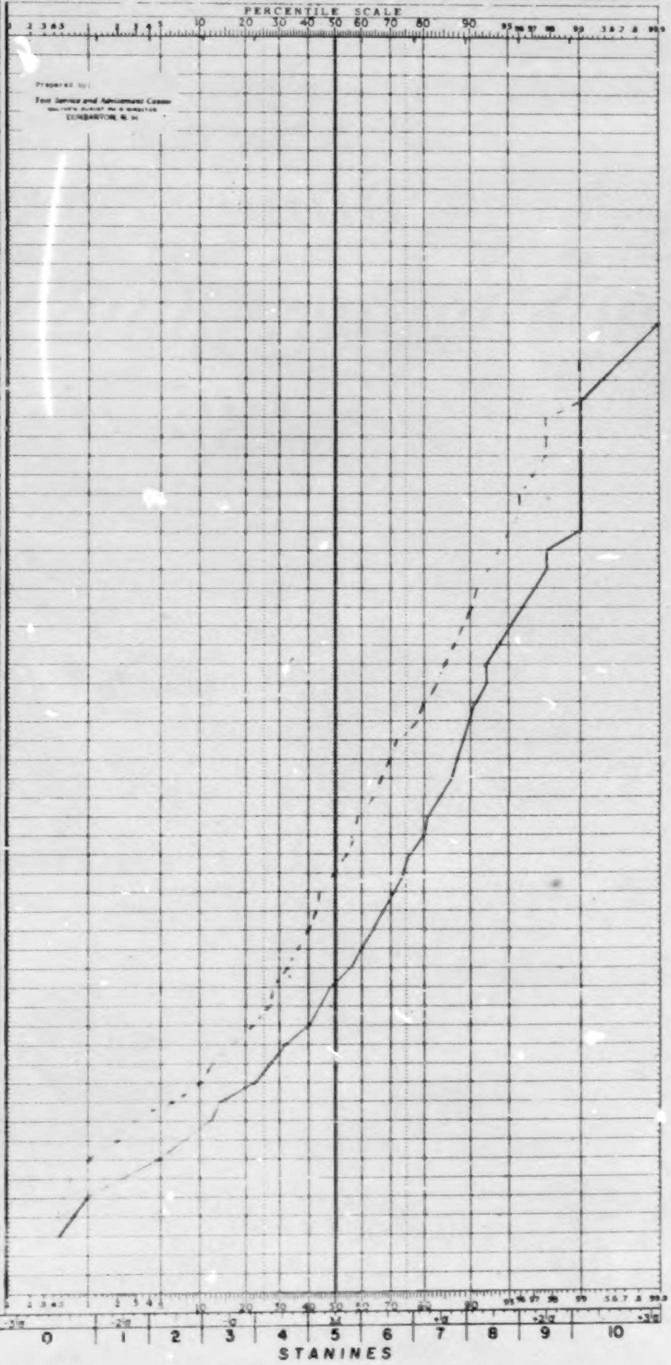
Plot of the Index of Guessing
Showing Decrease in the Amount of Guessing in the Spring Compared to Fall

TITLE I

NORMAL PERCENTILE CHART

	STUDY	Grade or Group	No. of Cases	MEASURE	Form	DATE	Graphs by	SCHOOL	COMMUNITY
I	Rights/Attempts-Word Meaning	4	422	SAT: Int. I	X	Fall '69	TS&AC	N.H. Statewide	Title I
II	" " _ " "	"	"	" " "	"	Spr. '70	"	" "	" "

SCORE INTERVALS					I Fall			II Spr.		
			*		Freq.	Per-cent	Cum. %	Freq.	Per-cent	Cum. %
650-	325-	195-	130-	65						
640-	320-	192	128-	64						
630-	315-	189-	126-	63						
620-	310-	186-	124-	62						
610-	305-	183-	122-	61						
600-	300-	180-	120-	60						
590-	295-	177-	118-	59						
580-	290-	174-	116-	58						
570-	285-	171-	114-	57						
560-	280-	168-	112-	56						
550-	275-	165-	110-	55						
540-	270-	162-	108-	54						
530-	265-	159-	106-	53						
520-	260-	156-	104-	52						
510-	255-	153-	102-	51						
500-	250-	150-	1.00-	50	2		100	3		100
490-	245-	147-	.98-	49	0		-	0		-
480-	240-	144-	.96-	48	0		-	1		99
470-	235-	141-	.94-	47	0		-	1		99
460-	230-	138-	.92-	46	1		99	2		99
450-	225-	135-	.90-	45	0		-	1		98
440-	220-	132-	.88-	44	0		-	2		98
430-	215-	129-	.86-	43	0		-	2		98
420-	210-	126-	.84-	42	1		99	3		97
410-	205-	123-	.82-	41	0		-	3		96
400-	200-	120-	.80-	40	2		99	2		96
390-	195-	117-	.78-	39	2		99	2		95
380-	190-	114-	.76-	38	2		98	8		94
370-	185-	111-	.74-	37	3		98	8		93
360-	180-	108-	.72-	36	2		97	3		91
350-	175-	105-	.70-	35	4		96	4		90
340-	170-	102-	.68-	34	4		95	10		89
330-	165-	99-	.66-	33	5		94	8		87
320-	160-	96-	.64-	32	3		93	9		85
310-	155-	93-	.62-	31	6		93	11		83
300-	150-	90-	.60-	30	7		91	10		80
290-	145-	87-	.58-	29	8		90	20		78
280-	140-	84-	.56-	28	4		88	13		73
270-	135-	81-	.54-	27	4		87	13		70
260-	130-	78-	.52-	26	9		86	18		67
250-	125-	75-	.50-	25	12		84	15		63
240-	120-	72-	.48-	24	2		81	6		59
230-	115-	69-	.46-	23	17		80	18		58
220-	110-	66-	.44-	22	10		[illegible]	16		53
210-	105-	63-	.42-	21	12		74	18		50
200-	100-	60-	.40-	20	15		71	6		45
190-	95-	57-	.38-	19	21		68	17		44
180-	90-	54-	.36-	18	12		63	12		40
170-	85-	51-	.34-	17	16		60	14		37
160-	80-	48-	.32-	16	33		56	22	-	34
150-	75-	45-	.30-	15	13		48	7		28
140-	70-	42-	.28-	14	22		45	14		27
130-	65-	39-	.26-	13	31		40	26		23
120-	60-	36-	.24-	12	26		32	24		17
110-	55-	33-	.22-	11	16		26	7		12
100-	50-	30-	.20-	10	34		23	16		10
90-	45-	27-	.18-	9	12		14	12		6
80-	40-	24-	.16-	8	18		12	7		3
70-	35-	21-	.14-	7	9		7	2		2
60-	30-	18-	.12-	6	13		5	2		1
50-	25-	15-	.10-	5	5		2	1		.7
40-	20-	12-	.08-	4	1		1	0		-
30-	15-	9-	.06-	3	1		.7	2		.5
20-	10-	6-	.04-	2	2		.5	0		-
10-	5-	3-	.02-	1	0		-	0		-
0-	0-	0-	.00-	0	0		-	0		-
MEDIANS					.31			.43		
STANDARD DEVIATION SCALE										



*Graphs are plotted on the step interval checked at the top of the chart.

THE USE OF ITEM DIFFICULTIES FOR PREDICTING THE TOTAL SCORE ON A TEST

In Part II we spent considerable time developing the history of a formula (which was first used back in the '50s in a generalized form) for predicting an untimed score from a time-limited score.

According to this formula, the last twenty items attempted by the individual in question was used, and the difficulty values of these items constituted the bottom of a ratio, the top of which was the number of items right on these twenty items.

The formula read: A + (B/C x D), where
A was the score on the test,
B was the score earned on the last twenty items attempted,
C was the sum of the difficulty values of those items,
D was the similar value for the remaining items in the test coming immediately subsequent to the last 20 items attempted.

We then adopted a modified special purpose formula; this time the formula always uses the first 20 items (23 in Paragraph Meaning to accommodate a paragraph break).

The special formula is more accurately written: A + (A'/C x D), where
A is the score made on the first 20 items,
A' is the same score,
C is the sum of the difficulty values of those twenty items,
D is the sum of the difficulty values of the remaining items.

For example, in Word Meaning there are 38 items; the first 20 are scored and the prediction is based upon the score earned in these 20 items according to the formula above.

Recall our basic assumption underlying this special purpose formula; namely, that we feel sure we will find a minimal amount of guessing in the first twenty items because these are the easiest items in the test.

A child would, and does, choose to answer the items he knows and marks his sheet according to his knowledge and not randomly; therefore, there should be a differential between the amount of guessing in the first 20 items and the last 18 (in the case of Word Meaning, for example).

We discovered in Part II that this procedure was not quite as simple to interpret as it seemed, and our hope that this might be a way of predicting a truer score for an individual than one he would get if he were allowed to guess randomly, especially toward the last part of the test, faded somewhat. It still is a most revealing type of analysis, especially for the Title I group.

Is the predicted score more valid with this group than the actual score? This may or may not be true. Without trying it out against some criterion, it would be very difficult to say.

We have carried out the same analytical procedure with the Title I population as for the random sample. Two tables are presented. In the first of these the means, standard deviations, and the numbers of cases are given separately for those children who Attempted-All items, those that Did-NOT-Attempt-All items, and for the Total Group. (See Table III-10.)

In a separate table, Table III-11, the correlations between the predicted score and the "number right" score earned on the test are given.

Perhaps we should consider first the table of means. All of these run lower, of course, than similar values for children in the random sample; for the most part the actual or "number right" means run slightly higher for the group of individuals who Attempted-All items than the mean of the predicted scores; which is what we expected.

In the case of those who Did-NOT-Attempt-All items, on the other hand, the mean scores tend to run slightly higher for the predicted score than the actually earned score.

This also makes sense because what we are doing in effect is giving credit for items the child has omitted, and sometimes a substantial number of items toward the end of the test are omitted because of their difficulty, rather than being answered randomly. These are the literal minded children; namely, those who do not feel threatened by the test.

A careful study of this table will reveal these trends, although it must be recalled to your mind that these are mean raw scores test-by-test and they can be compared directly, from group-to-group, only for a single test at a time.

In other words, the mean of the boys' group can be compared with the mean of girls for Word Meaning, and the mean of the predicted score for boys can be compared to the predicted score for girls for Computation, and so forth. You cannot compare the results for Word Meaning with Arithmetic Computation, etc.

Table III-10

Comparison of Attempted-All versus Did-NOT-Attempt-All Groups for Selected Statistics With Particular Emphasis on Magnitude and Direction of Differences Between Means

TITLE I

	Attempted-All			Did-NOT-Att.-All			Diff.of	Total Group*			Diff.of
	N	Mean	S.D.	N	Mean	S.D.	Means1/	N	Mean	S.D.	Means2/
Word Meaning											
BOYS											
Fall - Actual Score	59	11.4	4.8	199	8.3	4.6	+3.1	258	9.0	4.8	+2.4
Predicted Score		8.9	5.1		9.1	4.8	- .2		9.0	4.9	- .1
Spring-Actual Score	90	16.0	6.8	168	11.7	5.4	+4.3	258	13.2	6.3	+2.8
Predicted Score		14.8	7.0		13.9	6.4	+ .9		14.3	6.6	+ .5
GIRLS											
Fall - Actual Score	50	10.3	4.5	114	8.3	4.7	+2.0	164	8.9	4.7	+1.4
Predicted Score		7.8	3.9		9.3	5.0	-1.5		8.8	4.8	-1.0
Spring-Actual Score	56	13.6	6.0	107	12.8	5.6	+ .6	163	13.1	5.7	+ .5
Predicted Score		12.7	5.7		14.7	6.1	-2.0		14.0	6.1	-1.3
Paragraph Meaning											
BOYS											
Fall - Actual Score	83	17.6	4.7	177	15.2	6.4	+2.4	260	15.9	6.0	+1.7
Predicted Score		13.0	6.1		16.5	7.4	-3.5		15.3	7.2	-2.3
Spring-Actual Score	111	21.3	8.9	149	20.1	7.7	+1.2	260	20.6	8.2	+ .7
Predicted Score		20.6	9.8		24.4	9.8	-3.8		22.8	10.0	-2.2
GIRLS											
Fall - Actual Score	62	17.5	5.0	103	16.3	5.8	+1.2	165	16.7	5.6	+ .8
Predicted Score		13.3	5.0		18.2	6.7	-4.9		16.4	6.6	-3.1
Spring-Actual Score	73	21.5	6.9	90	20.7	7.4	+ .8	163	21.1	7.2	+ .4
Predicted Score		19.6	7.6		24.3	8.1	-4.7		22.2	8.2	-2.6
Arithmetic Computation											
BOYS											
Fall - Actual Score	104	10.3	3.8	156	9.8	4.0	+ .5	260	10.0	4.0	+ .3
Predicted Score		8.0	3.8		9.9	4.5	-1.9		9.1	4.3	-1.1
Spring-Actual Score	117	14.7	6.8	142	13.3	6.0	+1.4	259	13.9	6.4	+ .8
Predicted Score		13.2	6.2		14.5	5.9	-1.3		14.2	6.1	-1.0
GIRLS											
Fall - Actual Score	52	10.3	4.6	111	9.8	4.0	+ .5	163	9.9	4.2	+ .4
Predicted Score		8.3	4.3		10.3	4.2	-2.0		9.6	4.3	-1.3
Spring-Actual Score	65	15.1	6.7	98	14.8	5.7	+ .3	163	14.9	6.2	+ .2
Predicted Score		14.5	6.4		16.5	5.3	-2.0		15.7	5.9	-1.2

* Bivariate Overlays showing displacement of "Attempted-All" group versus "Did-NOT-Attempt-All" group shown in Appendix for Word Meaning and Arithmetic Comp.

1/ "Attempted-All" group versus "Did-NOT-Attempt-All" group

2/ "Attempted-All" group versus "Total Group"

Table III-10 - Attempted-All versus Did-NOT-Attempt-All - page 2

	Attempted-All			Did-NOT-Att.-All			Diff.of	Total Group			Diff.of
	N	Mean	S.D.	N	Mean	S.D.	Means1/	N	Mean	S.D.	Means2/
Arithmetic Concepts											
BOYS											
Fall - Actual Score	160	10.5	4.3	101	8.0	4.3	+2.5	261	9.5	4.5	+1.0
Predicted Score		10.1	4.5		8.9	4.4	+1.2		9.7	4.5	+ .4
Spring-Actual Score	190	12.7	5.1	68	10.7	5.4	+2.0	258	12.2	5.2	+ .5
Predicted Score		12.9	5.2		11.9	5.5	+1.0		12.7	5.3	+ .2
GIRLS											
Fall - Actual Score	99	9.5	4.1	66	8.1	4.0	+1.4	165	8.9	4.1	+ .6
Predicted Score		8.9	4.0		8.8	4.2	+ .1		8.9	4.1	0
Spring-Actual Score	114	11.5	5.0	51	10.1	4.6	+1.4	165	11.1	4.9	+ .4
Predicted Score		11.5	5.4		11.7	5.2	- .2		11.6	5.4	- .1
Arithmetic Applications											
BOYS											
Fall - Actual Score	169	9.3	4.3	91	8.2	4.4	+1.1	260	8.9	4.4	+ .4
Predicted Score		7.2	4.6		8.8	5.0	-1.6		8.6	4.7	-1.4
Spring-Actual Score	186	11.3	6.0	70	9.8	5.3	+1.5	256	10.9	5.9	+ .4
Predicted Score		11.1	6.4		10.8	6.1	+ .3		11.0	6.3	+ .1
GIRLS											
Fall - Actual Score	93	8.9	3.8	72	8.1	3.8	+ .8	165	8.6	3.8	+ .3
Predicted Score		8.2	3.7		8.7	4.0	- .5		8.4	3.8	- .2
Spring-Actual Score	115	10.7	5.4	50	10.7	4.4	0	165	10.7	5.1	0
Predicted Score		10.6	5.9		11.5	5.0	- .9		10.9	5.6	- .3

For any test, the mean of the Total Group is a kind of average of the two, weighted more heavily by the mean of the Did-NOT-Attempt-All group since the number of cases in this classification tends to be much larger.

This table of means and standard deviations is interesting in showing general trends, but not without the vexing exceptions reflecting the impurity of our hypothesis that the Did-NOT-Attempt-All group is free of guessing, etc.

The evidence presented by the investigation of the predicted score may not in itself be conclusive with respect to guessing in every individual case; it certainly is the basis for a strong presumption of substantive guessing if the predicted score is much lower than the earned score.

Individual variations in the predicted score versus the actual score suggests inconsistent behavior and (as was suggested in the similar table for random sample) the predicted score, in all probability, is the more valid estimate of what a child knows, as measured by this test, than the actual score reported from the scoring process for those children who earned low scores on the very easiest part of the test; i.e., the first twenty (23) items.

Item difficulties based upon large numbers of cases essentially constitute norms for these items. The prediction of total score from such data certainly is as persuasive evidence of guessing as a split-half reliability coefficient is evidence of the dependability of a test.

The proportion of children who increase their score very noticeably in the case of the predicted score is greater in inverse relationship to the magnitude of the raw score. How else than by guessing can inconsistency with the norms be explained?

Reference is made here to Table III-11, which provides further insight into this issue in the form of the correlation between actual and predicted scores. While these correlations are positive and high, the relationship is far from 1 to 1, as indicated by the coefficient of alienation.

The actual correlation plots are reproduced for two tests in Appendix B-5 - 8; namely, Word Meaning and Arithmetic Computa-

Table III-11

Correlation Coefficients and Coefficients of Alienation

Actual Scores versus Predicted Scores

TITLE I

	Attempted-All				Did-NOT-Att.-All				Total Group			
	Fall		Spring		Fall		Spring		Fall		Spring	
	r	k	r	k	r	k	r	k	r	k	r	k
Word Meaning												
Boys	.87	.49	.93	.37	.91	.41	.91	.41	.87	.49	.89	.46
Girls	.85	.53	.93	.37	.90	.44	.89	.46	.84	.54	.88	.47
Paragraph Meaning												
Boys	.81	.59	.90	.41	.72	.69	.79	.61	.66	.75	.81	.59
Girls	.79	.61	.85	.53	.79	.61	.77	.64	.69	.72	.75	.66
Arithmetic Computation												
Boys	.81	.59	.92	.39	.90	.44	.91	.41	.83	.56	.89	.46
Girls	.91	.41	.90	.44	.91	.41	.90	.44	.87	.49	.88	.47
Arithmetic Concepts												
Boys	.91	.41	.94	.34	.92	.39	.93	.37	.91	.41	.93	.37
Girls	.93	.37	.94	.34	.89	.46	.92	.39	.90	.44	.92	.39
Arithmetic Applications												
Boys	.93	.37	.96	.28	.93	.37	.95	.31	.92	.39	.96	.28
Girls	.92	.39	.95	.31	.92	.39	.93	.37	.90	.44	.95	.31

tion. These are in the form of bivariates in which the Attempted-All group is superimposed on the Did-NOT-Attempt-All group and is printed in a contrasting color to show the extent to which the Attempted-All group plot differs.

Those whom we have labeled generally as guessers, i.e. the Attempted-All'ers, tend to improve their score by guessing. The fact that they do so is indicated clearly by the fact that the distribution in color moves to the right; that is, in the direction of a higher score for the actual score than for the predicted score.

Statistics, it has been said, can be made to lie; but this corruption of the truth is done by the interpreter, not by the statistics if all the facts are given.

This writer's interpretation of these bivariate distributions is that there is, indeed, a real gain in the actual score by virtue of the guessing factor and that this operates more strongly on the harder portions of the test.

(A separate study is underway to compare the R/A ratio separately for the earlier, easier items and the final, hard items - which should be revealing.)

The best way to verify these data (and all of the previous data, for that matter) is to do it by careful study of the actual student protocols as presented in the rosters. The problem is that these rosters for this study are voluminous.

An additional sentence or two about some of the difficulties in applying the A + (A'/C x D) formula may be of interest to those who do not trust the final statistical summary - correlation or what not - bless them!

For example, some of those who marked all of the items actually earned a perfect or near-perfect score. For them, there was little reason to do any predicting at all. In fact, for the top 23% of the group the predicted score is of little value; i.e., stanines 7, 8, and 9.

This prediction procedure needs to be used in relation to the Rights/Attempts ratio discussed earlier and applies best in the case of those children who have a very low ratio. These children really and truly answered fewer right from sure knowledge in proportion to the number of items they have attempted, thus increasing the amount they benefitted (if you can call it that) by their guessing.

Perhaps these two techniques taken together constitute a valid way of identifying individuals who have guessed a great deal and, therefore, should be investigated carefully and possibly have their scores totally invalidated, depending on the total "profile" to wit:

1. A large proportion of easy items at the start of the test answered correctly.

2. Sometimes a gradual change and sometimes a sharp change to a guessing pattern of response.

3. Sometimes the "honest" child upsets the prediction by stopping abruptly and not using what partial knowledge he has.

Unfortunately for the Title I group, there are quite a few youngsters whose scores should be thrown out immediately, because in effect they are not representative scores anyway, in the sense that the curriculum they really need is not adequately represented by the test.

Effect of the Standard Correction for Guessing

If we apply to the scores the more or less standard correction for guessing, which for four-choice items would be R-1/3W, we get still another slant at these data. This was done for the boys in the Title I group.

A distribution of corrected scores was made for Word Meaning to show exactly what happens when the correction for guessing is applied, and negative scores are retained in negative form. It was also done for Arithmetic Computation. (Table III-12)

For Word Meaning, a total of 114 out of the group would have zero scores or less, and for Arithmetic Computation this number is 85.

Since a zero score is an "undistributed" score (i.e., not a point on a scale of scores), a great many of the actual (i.e., "rights") scores are wiped out and the test effort is wasted.

Table III-12

Distribution of Corrected Scores, Fall and Spring, Which Would Have Been Reduced to Zero or Negative Values 1/

TITLE I - Boys

Corrected Scores	R-1/3W Word Meaning	R-1/4W Arith. Comp.
0	29	25
-1	36	24
-2	20	19
-3	11	10
-4	6	2
-5	2	4
-6	4	1
-7	4	
-8	2	
Total Zero or Negative Values	114	85
Number of Cases, Fall and Spring	519	519
% of Cases Zero or Negative Values	22%	16%

1/ The above phenomenon happens when the formula R-1/3W or R-1/4W is applied to "correct" earned scores where 1/3W or 1/4W is the same or greater than number of Rights.

What about the "honest" individuals? They did NOT guess. Should they be penalized?

We can only identify them from the R/A ratio and a careful inspection of the rosters for item responses.

The Non-Relationship Corresponding to r

Turning back now to the table of correlations between the actual scores versus the predicted scores (Table III-11), we have introduced the coefficient of alienation.

At the time we discussed the Rights versus Attempts, we quoted from one authority. Here we're going to quote a somewhat similar passage from a different authority, which perhaps has a little bit different slant on the problem.

Referring to the correlation ratio, Peters and Van Voorhis (1/) say:

> "*This ratio* $\sqrt{1-r^2}$ *Kelley has named the 'coefficient of alienation.' It furnishes a very fruitful way of looking at a coefficient of correlation and of passing judgment as to how high a correlation must be in order to be satisfactory. The student (i.e., the teacher, administrator, etc.) should apply this test to correlation coefficients of various sizes. He will find that where r equals .10, there remains 99.5 per cent of guess in a prediction based on it; the prediction has been improved only one-half of 1 per cent over pure chance. Where r is .80, 60 per cent of chance still remains; we are only 40 per cent better off than if we drew predicted scores out of a hat. Even where r is .95, there remains 31 per cent of the element of chance in predicting placemement of individuals.*
>
> "*It is obvious that for the safe placement of individuals very high correlation coefficients are required, much higher than those called 'high' by others.*"

Kelley said that the situation is greatly different in working with means, but we're not concerned with this problem at the moment - we are concerned with individuals.

The correlations in the table we are presently discussing are on the high side, but just how high is high in this context?

The amount by which the sheer chance factor must be eliminated for sensible use of the estimated score data is indicated by the inserted "k" values alongside the correlation coefficients as a standard procedure under certain circumstances.

In the case of Paragraph Meaning, where the correlations of actual versus predicted score tend to be substantially lower, for some reason, it is easily seen that the coefficients of alienation are too high for us to take the correlation coefficients too seriously.

Summarizing the Title I Data to This Point

We have now been talking about the bivariate distributions, but if we consider both these and the R/A ratios as going certainly along together with our ultimate recourse (the examination of the actual rosters showing how the individual passed, failed, or omitted items) we perhaps can arrive at a series of generalizations concerning the statistical part of Title I comparable to our discussion of the random sample.

A. Individuals with high actual scores, and therefore apparently little guessing, gain little or nothing by our using the predicted score in place of the actual earned score. Perhaps we should generalize this to say that anyone falling in the top three stanines should be given the score he earned, regardless of the guessing factor, as the most valid measurement for him.

B. Children with low scores and consequently a very large guessing potential, since there are many more items on which to guess, should be examined carefully considering all data available. Guessing is maximized when they mark all of the items - as has happened so frequently among low-scoring pupils in both the Title I group and the random sample.

We can say that individuals with low scores, where individuals attempt far more items than they are able to answer correctly, can benefit most by having the actual score replaced by the predicted score; i.e., a more valid score is substituted for an inflated score.

Better yet, both the actually earned score and the predicted score can be given to reveal the extent to which guessing probably has affected the score.

1/ Peters, Charles C. and Walter R. VanVoorhis. Statistical Procedures and Their Mathematic Bases. New York: McGraw-Hill, 1940, 115.

SUMMARY AND CONCLUSIONS

Before anything else is said, it should be made amply clear that the authors represent this study to be true only of this population and this battery of tests. We suspect that much which follows will be found true of other batteries in other places, but this calls for cross validation studies.

Of these things we are almost certain:

1. There is no easy way to make a valid test, relatively free from guessing, without radical changes
    - a) in the Directions (no guessing)
    - b) in the item types used (approaching the job sample ideal) and
    - c) a dedication of teachers and administrators to the proposition that testing is as much a part of instruction and evaluation as is the textbook, lecture, demonstration or other teaching device.
2. We are reasonably sure that our data is convincing evidence that currently published multiple choice survey tests, without modification, are not useful in before-after measurement.
3. We are even surer that no formula ever will be devised that will correct for guessing with anywhere near universality, given the variations among tests (including Directions), teachers, and pupils.
4. We see no substitute for a conscientious personal judgment based on all available evidence before a conclusion is reached concerning the goodness of pupil performance relative to his ability to perform under all known influences.
5. We do not despair of success in overcoming most of the obstacles and difficulties described. Indeed, we dedicate ourselves to a continuing pursuit of these elusive goals on all fronts.

Appendix

". . . WHICH LEAVES US, AS WE SEEK SLIM CLUES TO TEACHING SUCCESS, EXACTLY WHERE? BACK ON THE LOG WITH MARK HOPKINS, I GUESS, SWEEPING EDUCATIONAL THEORIES ARE DECONSECRATED. THE STUBBORN FACT REDISCOVERED IS THAT A GOOD TEACHER PROVIDES A GOOD EDUCATION. A TRUSTING, JUST TEACHER WHO HAS A THEORY ABOUT TEACHING AND UNDERSTANDS IT PROPERLY AND WHO IS OUTGOING ENOUGH TO ROMANCE OTHERS WITH LEARNING GETS LOYAL RESPONSE." *Arsene O. Boykin, "There's a Unicorn in the Classroom," Phi Delta Kappan, April, 1974.*

Raw Score	Cum. %	Sta-nine	Fre-quency	
52	99	9	1	•
51	99	9	2	••
50	99	9	0	
49	99	9	2	••
48	99	9	2	••
47	99	9	1	•
46	99	9	0	
45	99	9	3	•••
44	98	9	8	••••••••
43	97	9	7	•••••••
42	96	8	9	•••••••••
41	94	8	7	•••••••
40	93	8	6	••••••
39	92	8	2	••
38	91	8	4	••••
37	91	8	8	••••••••
36	89	7	19	•••••••••••••••••••
35	86	7	12	••••••••••••
34	84	7	9	•••••••••
33	83	7	15	•••••••••••••••
32	80	7	18	••••••••••••••••••
31	77	6	15	•••••••••••••••
30	74	6	17	•••••••••••••••••
29	71	6	23	•••••••••••••••••••••••
28	68	6	22	••••••••••••••••••••••
27	64	6	20	••••••••••••••••••••
26	60	5	20	••••••••••••••••••••
25	57	5	18	••••••••••••••••••
24	54	5	14	••••••••••••••
23	51	5	18	••••••••••••••••••
22	48	5	27	•••••••••••••••••••••••••••
21	44	5	29	•••••••••••••••••••••••••••••
20	39	4	24	••••••••••••••••••••••••
19	35	4	27	•••••••••••••••••••••••••••
18	30	4	32	••••••••••••••••••••••••••••••••
17	25	4	22	••••••••••••••••••••••
16	21	3	24	••••••••••••••••••••••••
15	17	3	22	••••••••••••••••••••••
14	13	3	13	•••••••••••••
13	11	2	15	•••••••••••••••
12	8	2	13	•••••••••••••
11	6	2	9	•••••••••
10	4	2	5	•••••
9	4	1	7	•••••••
8	2	1	3	•••
7	2	1	4	••••
6	1	1	2	••
5	1	1	3	•••
4	1	1	0	
3	1	1	0	
2	1	1	1	•
1	1	1	1	•
			585	

Appendix A1

Frequency Distribution, Cumulative Percent Distribution, and Stanines Plus Histogram Showing Shape of Raw Score Distribution Graphically

RANDOM SAMPLE - PARAGRAPH MEANING - FALL 1969

Mean - 24.44 **St.Dev.-9.43**

Raw Score	Cum. %	Stanine	Frequency	
57	99	9	1	*
56	99	9	1	*
55	99	9	1	*
54	99	9	3	***
53	99	9	6	******
52	98	9	1	*
51	98	9	5	*****
50	97	9	8	********
49	96	8	11	***********
48	94	8	8	********
47	92	8	6	******
46	91	8	12	************
45	89	7	13	*************
44	87	7	18	******************
43	84	7	17	*****************
42	81	7	15	***************
41	79	7	12	************
40	76	6	14	**************
39	74	6	15	***************
38	72	6	19	*******************
37	68	6	22	**********************
36	65	6	25	*************************
35	60	5	11	***********
34	58	5	17	*****************
33	56	5	17	*****************
32	53	5	19	*******************
31	49	5	19	*******************
30	46	5	23	***********************
29	42	4	26	**************************
28	38	4	13	*************
27	36	4	16	****************
26	33	4	20	********************
25	29	4	21	*********************
24	26	4	15	***************
23	23	3	19	*******************
22	20	3	9	*********
21	19	3	18	******************
20	16	3	10	**********
19	14	3	14	**************
18	11	3	10	**********
17	10	2	14	**************
16	7	2	12	************
15	5	2	8	********
14	4	1	7	*******
13	3	1	3	***
12	2	1	3	***
11	2	1	4	****
10	1	1	2	**
9	1	1	1	*
8	1	1	2	**
7	1	1	0	
6	1	1	0	
5	1	1	1	*
			587	

Appendix A2

Frequency Distribution, Cumulative Percent Distribution, and Stanines
Plus Histogram Showing Shape of Raw Score Distribution Graphically

RANDOM SAMPLE - PARAGRAPH MEANING - SPRING 1970

Mean - 31.97

St. Dev. -10.50

```
 Raw   Cum.  Sta-   Fre-
Score   %    nine  quency

 30    99     9      1   *
 29    99     9      1   *
 28    99     9      0
 27    99     9      2   **
 26    99     9      5   *****
 25    98     9      1   *
 24    98     9      8   ********
 23    97     9      3   ***
 22    96     9     18   ******************
 21    93     8     20   ********************
 20    90     7     14   **************
 19    88     7     19   *******************
 18    84     7     34   **********************************
 17    79     6     23   ***********************
 16    75     6     25   *************************
 15    70     6     24   ************************
 14    66     6     25   *************************
 13    62     5     49   *************************************************
 12    54     5     60   ************************************************************
 11    43     4     42   ******************************************
 10    36     4     42   ******************************************
  9    29     4     47   ***********************************************
  8    21     3     34   **********************************
  7    15     3     34   **********************************
  6    10     2     27   ***************************
  5     5     1     15   ***************
  4     2     1     11   ***********
  3     1     1      1   *
  2     1     1      1   *
  1     1     1      1   *
                   587
```

Appendix A3
Frequency Distribution, Cumulative Percent Distribution, and Stanines
Plus Histogram Showing Shape of Raw Score Distribution Grahpically

RANDOM SAMPLE - ARITHMETIC CONCEPTS - FALL 1969
Mean - 12.88 St.Dev. - 5.20

```
 Raw   Cum.  Sta-   Fre-
Score   %    nine  quency

 31    99     9      1   *
 30    99     9      5   *****
 29    99     9      8   ********
 28    98     9      6   ******
 27    97     9      6   ******
 26    96     8     18   ******************
 25    92     8     19   *******************
 24    89     7     21   *********************
 23    86     7     16   ****************
 22    83     7     20   ********************
 21    79     6     35   ***********************************
 20    73     6     28   ****************************
 19    69     6     33   *********************************
 18    63     6     32   ********************************
 17    57     5     43   *******************************************
 16    50     5     34   **********************************
 15    44     5     29   *****************************
 14    39     4     21   *********************
 13    36     4     29   *****************************
 12    31     4     37   *************************************
 11    24     3     24   ************************
 10    20     3     26   ****************
  9    16     3     34   **********************************
  8    10     2     17   *****************
  7     7     2     18   ******************
  6     4     1     13   *************
  5     2     1      4   ****
  4     1     1      3   ***
  3     1     1      2   **
                   582
```

Appendix A4
Frequency Distribution, Cumulative Percent Distribution, and Stanines
Plus Histogram Showing Shape of Raw Score Distribution Graphically

RANDOM SAMPLE - ARITHMETIC CONCEPTS - SPRING 1970
Mean - 16.37 St.Dev. - 6.06

Raw Score	Cum. %	Sta-nine	Fre-quency	
32	99	9	1	*
31	99	9	1	*
30	99	9	6	******
29	99	9	4	****
28	98	9	3	***
27	97	9	13	*************
26	95	8	16	****************
25	92	8	13	*************
24	90	7	22	**********************
23	86	7	26	**************************
22	82	7	29	*****************************
21	77	6	25	*************************
20	73	6	25	*************************
19	68	6	33	*********************************
18	63	6	31	*******************************
17	57	5	27	***************************
16	53	5	36	************************************
15	47	5	36	************************************
14	40	4	27	***************************
13	36	4	35	***********************************
12	30	4	24	************************
11	26	4	22	**********************
10	22	3	29	*****************************
9	17	3	26	**************************
8	13	2	23	***********************
7	9	2	20	********************
6	5	1	16	****************
5	2	1	6	******
4	1	1	7	*******
3	1	1	1	*
			583	

Appendix A6
Frequency Distribution, Cumulative Percent Distribution, and Stanines
Plus Histogram Showing Shape of Raw Score Distribution Graphically

RANDOM SAMPLE - ARITHMETIC APPLICATIONS - SPRING 1970

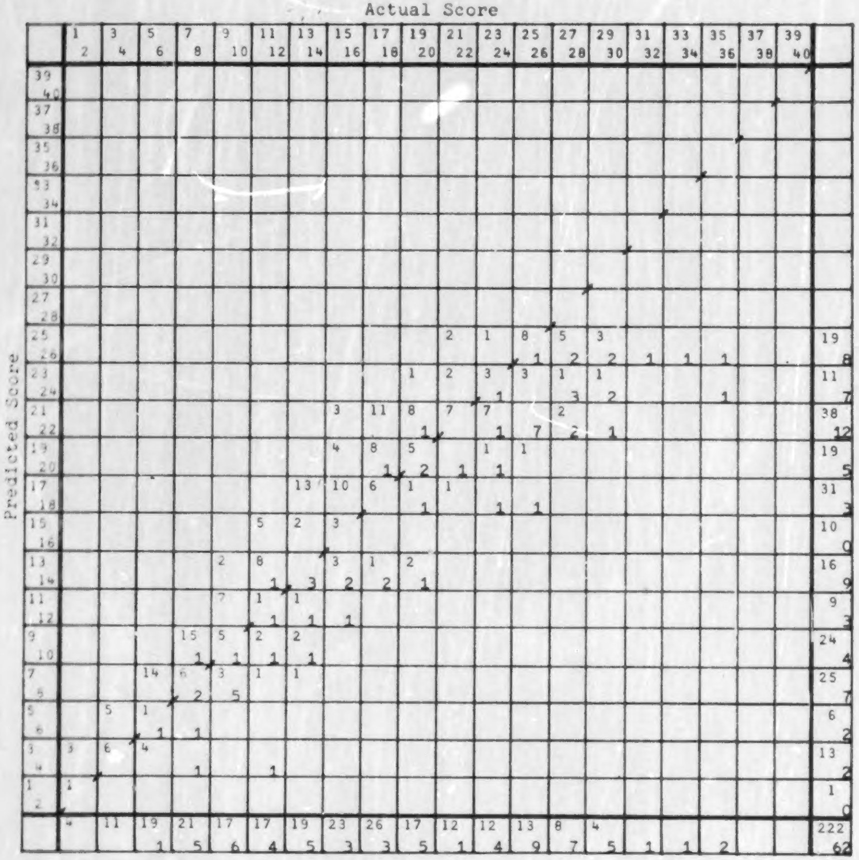
Mean - 16.21 St.Dev. - 6.21

Appendix B1 - Answer Sheet Study

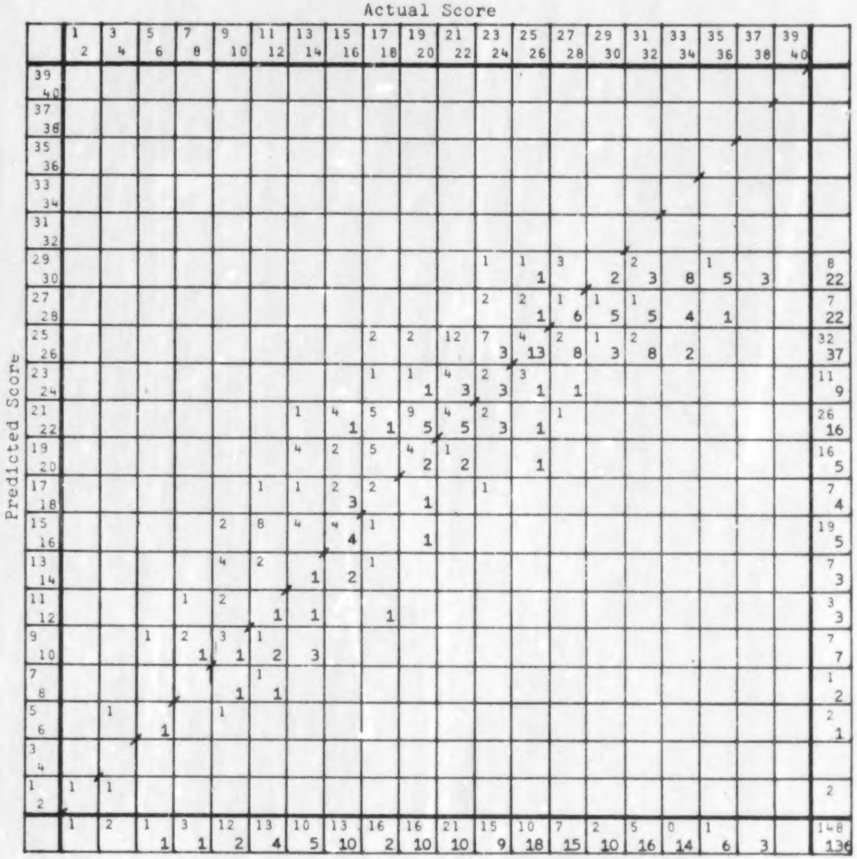
SUPERIMPOSED CORRELATION PLOTS FOR PREDICTED AND ACTUAL SCORES SEPARATELY FOR THOSE WHO ATTEMPTED ALL ITEMS AND FOR THOSE WHO DID NOT*

RANDOM SAMPLE - WORD MEANING - BOYS

FALL



SPRING



Note: Data derived from a Random Sample and Title I cases tested Fall & Spring 1969-70 with Stanford Achievement Test: Intermediate I: Form X.

* Colored figures in the lower right-hand corner of each cell are for those who attempted all items.

RANDOM SAMPLE - WORD MEANING - GIRLS

FALL

Actual Score

Predicted Score

	1 2	3 4	5 6	7 8	9 10	11 12	13 14	15 16	17 18	19 20	21 22	23 24	25 26	27 28	29 30	31 32	33 34	35 36	37 38	39 40	
39 40																					
37 38																					
35 36																					
33 34																					
31 32																					
29 30																					
27 28																					
25 26										2		6	3 1	4 3	2	 2	 1				17 7
23 24									1		6	4	2	1 2	1 2	 1					15 5
21 22								4	7	6	4	4 3	1 1	 4							26 8
19 20								10	5 2	5 2	3 1	 1	1 1	 1							24 8
17 18							10	7	7 1	5 4	 1										29 6
15 16						6	8	3 1	2 3												19 4
13 14					2	10	9 2	3 3	3 2	1											28 7
11 12					2	3	2	1 1													8 1
9 10				7	10 2	5 2	1 4														23 8
7 8				12 3	3 1	3	 1	 1													18 5
5 6		1	2	2 2																	5 2
3 4	3	2	1	1	 1																7 1
1 2																					0 0
	3	3	15	13 5	17 4	24 2	30 7	28 5	25 8	19 6	13 2	14 4	7 3	5 10	3 2	 3	 1				219 62

SPRING

Actual Score

Predicted Score

	1 2	3 4	5 6	7 8	9 10	11 12	13 14	15 16	17 18	19 20	21 22	23 24	25 26	27 28	29 30	31 32	33 34	35 36	37 38	39 40	
39 40																					
37 38																					
35 36																					
33 34																					
31 32																					
29 30														1	1 3	1 2	 4	 3	 1		3 13
27 28												1	1	2 3	1 4	1 7	1 3	 1			7 18
25 26									2	2 1	4 1	8 6	7 9	7 13	2 10	1 6	 2				33 48
23 24									1	7 1	3 3	7 3	1 4	 1	 1						19 13
21 22								5	7 3	5 8	10 2	3 5	1 1								31 19
19 20									5 3	3 2	1 1										9 6
17 18							2	3	3 1	2		 1									10 2
15 16					1	5	2 1	5 2	 2	1											14 5
13 14					3	2 1	1 2	 1													6 4
11 12				3	2		1 3														6 3
9 10					1		1														2 0
7 8			2		2																4 0
5 6			1		 1	 1															1 2
3 4		1	1																		2
1 2																					0
		1	4	3	9 1	7 2	7 6	13 3	18 9	20 12	18 7	19 15	10 14	10 17	4 18	3 15	1 9	 4	 1		147 133

Note: Data derived from a Random Sample and Title I cases tested Fall & Spring 1969-70 with Stanford Achievement Test: Intermediate I: Form X.

* Colored figures in the lower right-hand corner of each cell are for those who attempted all items.

RANDOM SAMPLE - ARITHMETIC COMPUTATION - BOYS

FALL

Actual Score

Predicted Score	1 2	3 4	5 6	7 8	9 10	11 12	13 14	15 16	17 18	19 20	21 22	23 24	25 26	27 28	29 30	31 32	33 34	35 36	37 38	39 40	
39 40																					
37 38																					
35 36																					
33 34																					
31 32																					
29 30																					
27 28																					
25 26																					
23 24																					
21 22									1	2	2		1								5 1
19 20								7	10 1	4 4	2	1	1								21 9
17 18								9	1												9 1
15 16							9 2	3 1	2 2	1											15 5
13 14						20	8 3	3 1													31 4
11 12					21 1	8 7	4 4	1 2	1												34 15
9 10				13	11 1	4 2	3														28 6
7 8			3 1	23 2	7 7	2 4	2														35 16
5 6		6	6 4	4 3	7 3	1															24 10
3 4		6 1	1 2	1																	7 4
1 2	2		2 1																		4 1
	2	12 1	12 8	40 6	46 12	35 13	21 14	23 4	13 5	7 4	2 2	1	2								213 72

Note: Data derived from a Random Sample and Title I cases tested
Fall & Spring 1969-70 with Stanford Achievement Test:
Intermediate I: Form X.

SPRING

Actual Score

Predicted Score	1 2	3 4	5 6	7 8	9 10	11 12	13 14	15 16	17 18	19 20	21 22	23 24	25 26	27 28	29 30	31 32	33 34	35 36	37 38	39 40	
39 40																					
37 38																					
35 36																					
33 34																					
31 32																					
29 30																					
27 28											1				1	1 1			2		3 3
25 26										1	2	1	2	5 1	1 3	2	1				12 7
23 24									3 1	4 1	7 1	7 4	2 2	4 1	2 2	3	1				29 16
21 22								3	4	5	6	1 1	2 2	1 1	1						22 5
19 20							1	13 1	7 3	6 3	5 2	4	3	1	1						32 18
17 18							6	1 1	1	3 4		1									11 6
15 16						6	5 3	2 6	2 5	3											15 17
13 14					1	3	2	3 2	1	1											7 6
11 12				3	7	3 5	3 3	1	2												17 10
9 10				3	2 1	1 3	1														6 5
7 9			2 1	5 3	4 4	1 5															12 13
5 6		1		1	1																2 1
3 4		1	3	2		1															2 5
1 2		1																			0 1
		2 1	2 4	12 5	14 6	15 13	15 9	23 10	17 12	19 12	21 3	9 10	6 7	10 4	4 7	1 6	2	0	2	0	170 113

* Colored figures in the lower right-hand corner of each cell are for those who attempted all items.

RANDOM SAMPLE - ARITHMETIC COMPUTATION - GIRLS

FALL

Actual Score

Predicted Score

	1 2	3 4	5 6	7 8	9 10	11 12	13 14	15 16	17 18	19 20	21 22	23 24	25 26	27 28	29 30	31 32	33 34	35 36	37 38	39 40	
39 40																					
37 38																					
35 36																					
33 34																					
31 32																					
29 30																					
27 28																					
25 26																					
23 24												1 1									1 1
21 22									 1	1	3	2 1			 1						6 3
19 20								3	9	5			1								18 0
17 18								4	4	 1	1 1										9 2
15 16							21	18 1	1 4	2 1											42 6
13 14					1	20 4	4 2	3 2	1												29 8
11 12					29 1	7 4	2 3	 2													38 10
9 10				11	6 2	1 3	2 1														20 6
7 8			13	13 2	6 6	3 5															35 13
5 6		5	9 1	1 2	2 3	 1															17 7
3 4		2	3 1	 2																	5 3
1 2	1																				1 0
	1	7	25 2	25 6	44 12	31 17	29 6	28 5	15 5	8 2	4 1	3 2	1 0	 0	 1						221 59

Note: Data derived from a Random Sample and Title I cases tested Fall & Spring 1969-70 with Stanford Achievement Test: Intermediate I: Form X.

SPRING

Actual Score

Predicted Score

	1 2	3 4	5 6	7 8	9 10	11 12	13 14	15 16	17 18	19 20	21 22	23 24	25 26	27 28	29 30	31 32	33 34	35 36	37 38	39 40	
39 40																					
37 38																					
35 36																					
33 34																					
31 32																					
29 30																					
27 28												2	 1	1 1	1	2 1	1	 1			7 4
25 26										1	2	2 1	1	6 3	5	2	 3	 1			19 8
23 24									4 1	10	13 1	7 6	3 4	6	6 5	1	 1				50 18
21 22								2	4	5	3 4	2 1	1 2		1						18 7
19 20							 1	12 1	6 4	3 5	3 6	1 2	 1	 1							25 21
17 18							7 1	2 1	2 1	3 2											14 [illegible]
15 16						8 1	5 6	5 2													18 [illegible]
13 14					1	1 3	3 3	 1	 2												5 9
11 12				3	4	4 4	3 3	 2													14 9
9 10				3 1	2																5 1
7 8			2 1	5		1 1															8 2
5 6			1	1																	2 0
3 4		1	1 1																		2 1
1 2																					0 0
		1	4 2	12 1	7 0	14 9	18 14	21 7	16 8	22 7	21 11	14 10	5 8	13 5	13 5	5 1	1 4	 2			187 94

* Colored figures in the lower right-hand corner of each cell are for those who attempted all items.

TITLE I - WORD MEANING - BOYS

FALL

Actual Score

Predicted Score	1 2	3 4	5 6	7 8	9 10	11 12	13 14	15 16	17 18	19 20	21 22	23 24	25 26	27 28	29 30	31 32	33 34	35 36	37 38	39 40	
39 40																					
37 38																					
35 36																					
33 34																					
31 32																					
29 30																					
27 28																					
25 26										1			1			1					2 1
23 24																					0 0
21 22								1	[illegible] 1			2									4 1
19 20									1 1		1										2 1
17 18							5	2	3	1 3	1										11 4
15 16						2	3	1 1		1											6 2
13 14					3	9 2	3 1	2	1 1												18 4
11 12					7	6	2 1														15 1
9 10				20 2	15 4	2 2	1 3														38 11
7 8			30	11 2	2 6	1 6	1														44 15
5 6		10	6	3 3	1 3	1	1														21 7
3 4	6	20	4 3	1 3	2	2															31 10
1 2	4	1 2	2																		7 2
	10	31 2	40 3	37 10	28 15	21 12	14 7	6 1	6 3	2 4	1 1	2 0	1 0	0	0	1					199 59

SPRING

Actual Score

Predicted Score	1 2	3 4	5 6	7 8	9 10	11 12	13 14	15 16	17 18	19 20	21 22	23 24	25 26	27 28	29 30	31 32	33 34	35 36	37 38	39 40	
39 40																					
37 38																					
35 36																					
33 34																					
31 32																					
29 30																					
27 28											1		1		3		2				1 6
25 26									[illegible]	2	1	1	1		1						5 4
23 24									2	3 1	3	1 3		1							9 5
21 22							1	6	5	4 1	4 6	2		1							20 10
19 20							7	1 1	2 2												10 3
17 18						4	1	8 3	4 3												17 6
15 16					1	6 1	7 4	6 1	3	2 2	1										22 12
13 14					6 1	3	4 2	2	2												13 7
11 12				3	3	1 1	2	1													8 3
9 10			9	9	7 5	1 3	4	2													26 14
7 8			10	2	3 7																13 9
5 6		6	3 1	1 2	2	1															10 6
3 4	2	4 1	3 1	2 2	1 1																12 5
1 2	2																				2
	4	10 1	25 2	15 6	21 16	15 6	20 12	22 9	16 10	11 4	8 8	1 6	2	2	4	0	2				168 90

Note: Data derived from a Random Sample and Title I cases tested Fall & Spring 1969-70 with Stanford Achievement Test: Intermediate I: Form X.

* Colored figures in the lower right-hand corner of each cell are for those who attempted all items.

TITLE I - WORD MEANING - GIRLS

FALL

Actual Score

Predicted Score

	1 2	3 4	5 6	7 8	9 10	11 12	13 14	15 16	17 18	19 20	21 22	23 24	25 26	27 28	29 30	31 32	33 34	35 36	37 38	39 40	
39 40																					
37 38																					
35 36																					
33 34																					
31 32																					
29 30																					
27 28																					
25 26																					
23 24									1		1				 1						2 1
21 22										1											1 0
19 20									2		1										3 0
17 18							5	3	1	1		 1									10 1
15 16						1	1 1		1 1												3 2
13 14			1		2	4	1		 1												8 1
11 12					5	2															7 0
9 10				10	7 5	1 2	1 2	1 1													20 10
7 8			17	4 9	4 4	 5	1	 1													26 19
5 6		8	3	3 3	 1	 2															14 6
3 4	3	11 1	1 4	 2	1 1	 1															16 9
1 2	4		 1																		4 1
	7	19 1	22 5	17 14	19 11	8 10	9 3	4 2	5 2	2 0	2 0	 1	 0	 0	 1						114 50

Note: Data derived from a Random Sample and Title I cases tested Fall & Spring 1969-70 with Stanford Achievement Test: Intermediate I: Form X.

SPRING

Actual Score

Predicted Score

	1 2	3 4	5 6	7 8	9 10	11 12	13 14	15 16	17 18	19 20	21 22	23 24	25 26	27 28	29 30	31 32	33 34	35 36	37 38	39 40	
39 40																					
37 38																					
35 36																					
33 34																					
31 32																					
29 30																					
27 28												1	1			1					3
25 26									1	2	1		 1			 2					4 3
23 24									1	1	1 1	1	1 1								5 2
21 22							2	2	2	1	2 1	1	 1								10 2
19 20							1	5 1	2 1	 1											8 3
17 18								3	2 2	1 1											6 3
15 16					2	9	7 1	3 3	2 2												23 6
13 14					3	1 1	3 3	 1	2												9 5
11 12				3	4 1	1 1		 2													8 4
9 10			4	5 2	4 11	 1	1 2														14 16
7 8			2 1	3 1	 4	 1	 1														5 8
5 6		2	1 1	3 1																	6 2
3 4		2	2 1	1 1	1																6 2
1 2																					0 0
		4	9 3	15 5	14 16	11 4	14 7	13 7	12 5	5 2	4 2	3 0	2 3	0 0	0 0	1 2					107 56

* Colored figures in the lower right-hand corner of each cell are for those who attempted all items.

TITLE I - ARITHMETIC COMPUTATION - BOYS

FALL

Actual Score

Predicted Score

	1 2	3 4	5 6	7 8	9 10	11 12	13 14	15 16	17 18	19 20	21 22	23 24	25 26	27 28	29 30	31 32	33 34	35 36	37 38	39 40	
39 40																					
37 38																					
35 36																					
33 34																					
31 32																					
29 30																					
27 28																					
25 26																					
23 24																					
21 22										3											3
19 20									1 1												1 1
17 18								2	2	2 1											6 1
15 16							9	6 2	2 1	 2											17 5
13 14						13 2	3 1	3 2	1 2												20 7
11 12					14 2	6 3	2 3	 4	 1												22 13
9 10				7	 4	5 3	1 1														13 8
7 8			10	18 9	8 9	1 4	 4	 1													37 27
5 6		5 1	10 6	6 6	2 8	 4	 2														23 27
3 4		3	3 6	 3	1 1																7 10
1 2	1	 2	5 2	1 1																	7 5
	1	8 3	28 14	32 19	25 24	25 16	15 11	11 9	6 5	5 3											156 104

Note: Data derived from a Random Sample and Title I cases tested Fall & Spring 1969-70 with Stanford Achievement Test: Intermediate I: Form X.

SPRING

Actual Score

Predicted Score

	1 2	3 4	5 6	7 8	9 10	11 12	13 14	15 16	17 18	19 20	21 22	23 24	25 26	27 28	29 30	31 32	33 34	35 36	37 38	39 40	
39 40																					
37 38																					
35 36																					
33 34																					
31 32																					
29 30																					
27 28														 1	1						1 1
25 26											 1	2			 1	 1		 1			2 4
23 24									1	2	2 1	 1	1 2	3 1	 2						9 7
21 22									4	5 1	3	2 1		1	 2						15 4
19 20							3	5 3	7 1	3 5	1 1	 2	1 1								20 13
17 18							5 1	3 2	 1	 1		 1									8 6
15 16						4	8 3	1 6	 4	1 2		 1									14 16
13 14					4	4	5 1	1 3													14 4
11 12			1	3	9 3	4 6	4 9	 2													21 20
9 10				9 1	2 1	 3	1														12 5
7 8			6 1	3 8	5 8	1 4	1 1	 1													16 23
5 6		2	 2	1 3		 1															3 6
3 4	1	3 3	 3	1 1	 1																5 8
1 2	2																				2 0
	3	5 3	7 6	17 13	20 13	13 14	27 15	10 17	12 6	11 9	6 3	4 6	2 3	4 2	1 5	 1	 0	 1			142 117

* Colored figures in the lower right-hand corner of each cell are for those who attempted all items.

TITLE I - ARITHMETIC COMPUTATION - GIRLS

FALL

Actual Score

Predicted Score	1 2	3 4	5 6	7 8	9 10	11 12	13 14	15 16	17 18	19 20	21 22	23 24	25 26	27 28	29 30	31 32	33 34	35 36	37 38	39 40	
39 40																					
37 38																					
35 36																					
33 34																					
31 32																					
29 30																					
27 28																					
25 26																					
23 24																					
21 22									1		1										1 1
19 20								1	2												3 0
17 18									1	1 1											2 1
15 16							7	2	2 3	1	1										12 4
13 14						10	3 1	1	2												13 4
11 12					14 1	8 2	4 1	1 1													27 5
9 10				5	3	3 2															11 2
7 8			7 1	9 4	2 3	2	1 2														19 12
5 6		4	8 5	3 7	1 4																16 16
3 4		4	2	2	1																4 5
1 2	2	1 2																			3 2
	2	9 2	15 8	17 13	20 9	21 6	15 4	4 2	6 5	2 1	2										111 52

SPRING

Actual Score

Predicted Score	1 2	3 4	5 6	7 8	9 10	11 12	13 14	15 16	17 18	19 20	21 22	23 24	25 26	27 28	29 30	31 32	33 34	35 36	37 38	39 40	
39 40																					
37 38																					
35 36																					
33 34																					
31 32																					
29 30																					
27 28																	1		1		2
25 26												1		1	1						2 1
23 24									5	2 2	3 1	2 2	2	1 2	1	1					16 8
21 22								1	3	2 3	1 2										7 5
19 20							2	5 2	4	3	1		1								15 3
17 18							2 2	6	1 2	1											10 4
15 16						3	4 3	2 2	3 2	1											13 7
13 14					5	5 2	1 2	2													13 4
11 12				1	4 1	3 6	1	1													8 9
9 10				2 1	4	2 1	3														4 9
7 8			7	1 3	4	1															8 8
5 6		1	1	1		1															1 3
3 4			1	1																	0 2
1 2	1																				1 0
	1	1	7 2	4 6	9 9	13 11	9 11	16 5	16 4	9 5	5 3	2 3	2 1	2 2	2 0	1	1	0	1	0	98 65

Note: Data derived from a Random Sample and Title I cases tested
Fall & Spring 1969-70 with Stanford Achievement Test:
Intermediate I: Form X.

* Colored figures in the lower right-hand corner of each cell are for those who attempted all items.

APPENDIX C

Correlation Coefficients

Actual Scores versus Predicted Scores

RANDOM SAMPLE

	Attempted-All		Did-NOT-Att.-All		Total Group	
	Fall	Spring	Fall	Spring	Fall	Spring
Arithmetic Concepts						
Boys	.94	.95	.90	.91	.92	.94
Girls	.94	.95	.92	.93	.92	.94
Arithmetic Applications						
Boys	.94	.96	.94	.93	.93	.95
Girls	.92	.94	.94	.93	.92	.94

Comparison of Attempted-All versus Did-NOT-Attempt-All Groups for Selected Statistics

With Particular Emphasis on Magnitude and Direction of Differences Between Means

RANDOM SAMPLE

	Attempted-All			Did-NOT-Att.-All			Diff.of Means1/	Total Group			Diff.of Means2/
	N	Mean	S.D.	N	Mean	S.D.		N	Mean	S.D.	
Arithmetic Concepts											
BOYS											
Fall - Actual Score	183	13.9	5.6	101	12.5	4.6	+1.4	284	13.4	5.3	+ .5
Predicted Score		13.3	5.5		13.5	4.7	- .2		13.4	5.2	- .1
Spring-Actual Score	226	17.2	6.4	57	14.8	5.7	+2.4	283	16.7	6.3	+ .5
Predicted Score		17.0	6.3		16.0	6.0	+1.0		16.9	6.3	+ .1
GIRLS											
Fall - Actual Score	168	13.1	5.2	114	11.6	4.7	+1.5	282	12.5	5.1	+ .6
Predicted Score		12.5	5.2		12.4	4.7	+ .1		12.5	5.0	0
Spring-Actual Score	213	16.5	6.0	66	14.5	5.2	+2.0	279	16.0	5.9	+ .5
Predicted Score		16.4	5.9		15.2	5.0	+1.2		16.1	5.7	+ .3
Arithmetic Applications											
BOYS											
Fall - Actual Score	167	13.7	5.2	115	12.8	4.8	+ .9	282	13.3	5.1	+ .4
Predicted Score		13.0	5.3		13.7	5.2	- .7		13.3	5.2	- .3
Spring-Actual Score	218	16.9	6.9	66	16.0	5.0	+ .9	284	16.5	6.5	+ .4
Predicted Score		15.8	6.9		17.3	5.2	-1.5		16.4	6.6	- .6
GIRLS											
Fall - Actual Score	140	12.9	5.1	141	11.7	4.6	+1.2	281	12.3	4.9	+ .6
Predicted Score		12.4	5.0		12.5	4.6	- .1		12.5	4.8	- .1
Spring-Actual Score	200	16.4	5.8	79	14.6	5.9	+1.8	279	15 9	5.9	+ .5
Predicted Score		16.3	6.0		15.7	5.6	+ .6		16.1	5.9	+ .2

1/ "Attempted-All" group versus "Did-NOT-Attempt'All" group

2/ "Attempted-All" group versus "Total Group"